STRUCTURE AND FUNCTION OF LENTIVIRAL GENOMIC AND MESSENGER RNA

Elizabeth Grace Pollom

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biochemistry and Biophysics.

Chapel Hill
2012

Approved By:

Ronald Swanstrom, Ph.D.

Kevin M. Weeks, Ph.D.

William F. Marzluff, Ph.D.

Charles W. Carter Jr., Ph.D.

Howard M. Fried, Ph.D.

ABSTRACT

ELIZABETH GRACE POLLOM: Structure and Function of Lentiviral Genomic and Messenger RNA
(Under the direction of Ronald Swanstrom, Ph.D.)

The positive sense lentiviral RNA genome is packaged within the virus as a dimer of two single strands. The RNA of primate lentiviruses human immunodeficiency virus (HIV-1) and simian immunodeficiency virus (SIVmac239) are distantly related and the secondary structures of these viral RNAs share many known biological functions. Using selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE), I present an analysis of the secondary structure of *ex virio* genomic SIVmac239 RNA in relation to that of HIV-1 as well as an investigation into the secondary structure of the various *in vitro* mRNA species of HIV-1 resolved using the SHAPE technique.

First, I describe a SHAPE-derived model of SIVmac239 genomic RNA structure. When compared to that of HIV-1, I find very few conserved structural regions outside the previously studied functional structures. I observe that this is due to the flexible nature of the adenosine-rich lentiviral genome. The structures that are conserved are located in regions with high guanosine concentration, forming more stable pairing interactions. These results suggest that lentiviral genomic RNA structure is flexible and metastable unless held by stronger pairing interactions that seem to persist through the course of viral evolution.

The lentiviral genomic RNA structures that I have studied do share a few common base pairs, including a small stem-loop at the site of the first splice acceptor (SA1). In the second part, I describe the effect of mutating this structure on viral replication and on the splicing profile of the viral mRNA. To further investigate viral splicing regulation, I determined the SHAPE-derived structures of the most abundant mRNA variants for all of the protein products of HIV-1. Results reveal local interactions that form at regulatory regions in the viral transcripts.

Because RNA is an important feature throughout the lentiviral replication cycle, a greater understanding into the role of RNA in various aspects of viral replication will increase comprehension toward the complex biology of infection. This analysis provides insight into evolutionary conservation of RNA structures that play functional roles and may be possible targets for novel factors as part of a broad spectrum of viral inhibitory agents.

To my family: to Mom and Dad, Katie, Scotty, Anna, and Emily.

From your Carolina visits to my trips home for holidays, from the computer conversations to the late-night phone calls, from the sharing of music to the sharing of food, from the sacrifices you have made to the prayers you have said, from the tears to the laughter, from Indianapolis to West Lafayette to Bloomington to Boston to Chicago to St. Louis to London to Mumbai…

You have always been in North Carolina in my heart helping me along the way.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figures

ABBREVIATIONS

ΔG	change in Gibbs Free Energy
1M7	1-methyl-7-nitroisatoic anhydride
3'ss	3' splice site
^{33}P-αdCTP	^{33}P-labeled deoxycytidine triphosphate
5'ss	5' splice site
A	adenine
Å	angstrom
AIDS	acquired immunodeficiency syndrome
APOBEC	apolipoprotein B mRNA-editing, enyme-catalytic, polypeptide-like
b	intercept
bp	base pair
C	cytosine
CA	Capsid
cDNA	complementary DNA
cPPT	central polypurine tract
DIS	dimerization initiation site
DMEM	Dulbecco's modified Eagle medium
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
Env	Envelope
ESE	exonic splicing enhancer

ESS	exonic splicing silencer
G	guanine
G-P-P FS	Gag-Pro-Pol Frameshift
Gag	Gag polyprotein precursor
Gag-Pro-Pol	Gag-Protease-Polymerase polyprotein precursor
gp120	glycoprotein 120 (Env surface subunit)
group M	main group
group N	new group
group O	outlier group
HIV	human immunodeficiency virus
hnRNP	heterogenous nuclear ribonucleoprotein
hr	hours
IDT	Integrated DNA Technologies
IN	Integrase
ISE	intronic splicing enhancer
ISS	intronic splicing silencer
kb	kilobase
m	slope
M	molar
MA	Matrix
Matlab	Matrix laboratory
min	minutes
ml	milliliter

mM	millimolar
NC	Nucleocapsid
nt	nucleotide
PBS	primer binding site
PCR	polymerase chain reaction
pg41	glycoprotein 41 (Env transmembrane subunit)
pH	negative log of the hydrogen concentration
Pol II	cellular polymerase II
PPT	polypurine tract
PR	Protease
Pro-Pol	Protease-Polymerase polyprotein precursor
py-tract	polypyrimidine tract
R	repeat
RNA	ribonucleic acid
RNase	ribonuclease
RPMI	Roswell Park Memorial Institute medium
RRE	Rev-response element
rRNA	ribosomal RNA
RT	Reverse Transcriptase
SA	splice acceptor
SD	splice donor
sec	seconds
SHAPE	selective 2'-hydroxyl acylation analyzed by primer extension

siRNA	small interfereing RNA
SIV	simian immunodeficiency virus
SLS	stem-loop structure
SLSA1	stem-loop splice acceptor 1
snRNA	small nuclear RNA
snRNP	small nuclear RNP
SP	signal peptide
SR	serine arginine repeat protein
SRE	splicing regulatory element
SRP	signal recognition particle
ss	single stranded
TAR	trans-acting response
TBE	tris borate EDTA
TF	transition frame
tRNA	transfer RNA
U	uracil
U2AF	U2 activating factor
U3	unique 3' region
U5	unique 5' region
UTR	untranslated region
v/v	volume per volume
μCi	microcuries
μg	microgram

µl	microliters

CHAPTER I

LENTIVIRUSES AND RNA STRUCTURE

A. Basic biology of lentiviruses

1. Genome organization and viral replication

Human immunodeficiency virus (HIV) and simian immunodeficiency virus (SIV) belong to the family *retroviridae* in the genus lentivirus. Retroviruses are defined by their common replication strategy which involves a reverse-transcribed DNA intermediate that is integrated into the genomic DNA of the host cell and by virions composed of structural proteins surrounding and encapsidating viral enzymes that are packaged within the virion along with the two strands of single-stranded, positive sense RNA. The lentiviruses have additional accessory proteins that function as adaptor proteins by interacting with host factors in the cell (1).

The RNA genomes of these viruses largely represent coding domains for genes of a number of viral proteins (Figure 1.1). The intrinsic function of the genomic RNA is to be immediately reverse-transcribed into DNA, which is integrated into the cellular genome and eventually transcribed into full length genomic RNA by the cellular transcription machinery. This RNA has the baseline functions of acting as a template for protein translation, both in its full length form and as spliced forms, and packaging into new virions as genomic RNA (1). These basic functions, however, cannot fully account for the entirety of the RNA. The RNA sequence provides a context for replication and

translation, but RNA structure has implications in transcription activation, dimerization, packaging, frameshifting, nuclear export, and splicing regulation. Such functions are incompletely understood and require a more complete description to explain viral replication.

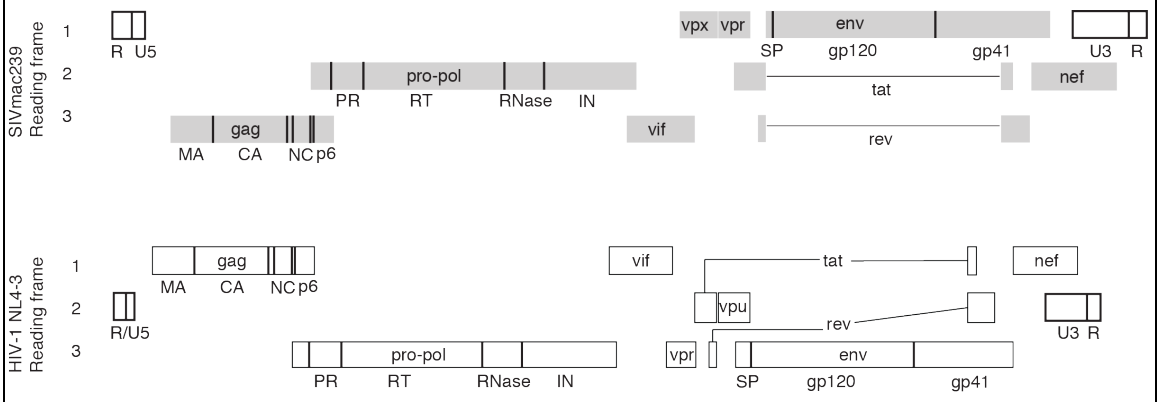


Figure 1.1: Organization of SIVmac239 and HIV-1 genomes. The SIVmac239 (above) protein-coding domains are in gray (labeled), and lines indicate domain junctions. The domains are not shown in codon alignment with the HIV-1_{NL4-3} genome (below) whose protein-coding domains shown in white.

The lentiviral replication cycle, including the function of many viral proteins, revolves around RNA. When the envelope protein (Env) interacts with the cellular receptor CD4 along with either the coreceptor CCR5 or CXCR4, the virus enters the host cell with a pair of identical, positive sense, single-stranded RNA genomes. Inside the cell, viral reverse transcriptase (RT) transcribes this approximately 9kb RNA into one copy of double-stranded DNA, then integrase (IN) incorporates the DNA into the host genome. The DNA is transcribed to produce genomic RNA, some of which is spliced to form the various mRNAs, which code for all of the structural and enzymatic viral proteins, as well as the accessory proteins Vpr and Vpu (in HIV), Vpx (in SIV), Vif, Nef, Tat, and Rev. The host-initiated transcription of the viral DNA produces full-length 5'-capped and 3'-polyadenylated RNA. The full-length RNA is packaged by and with the structural Gag and Gag-Pro-Pol polyproteins. The surrounding envelope is composed of cellular lipids and viral gp120 and gp41 glycoproteins, which are cleaved from the full-length Env protein by a cellular protease. Viral protease (PR) cleaves Gag and Gag-Pro-Pol into individual structural nucleocapsid (NC), matrix (MA), and capsid (CA), and the enzymes PR, RT and IN proteins, creating a virion that is mature and infectious (1).

2. Viral evolution

Two distinct types of HIV are currently circulating in the human population, and both are causative agents of AIDS. HIV type 1 (HIV-1) is the main type in most of the world, while HIV type 2 (HIV-2) is localized mainly to West Africa. Although both viruses use the same cellular receptors and co-receptors for entry, HIV-2 progresses to AIDS at a slower rate than HIV-1 and appears to be less infectious (77, 113). The

mutagenic rate of HIV-1 has propelled its genetic diversity via pressure from host immune system and restriction factors acting in conjunction with an error-prone reverse transcriptase and a short life cycle (43). Through this rapid evolution, the virus has been able to adapt to its host and be successful in high level replication as it has moved from its original hosts of non-human primates to humans (reviewed in (67)).

The lentiviruses found in non-human primates, SIV, infect African apes or Old World monkeys with little capacity to cause disease in these animals. One strain of SIV (SIVmac) originating from a species of West African monkeys, called sooty mangabey, is able to cause illness in a group of Asian monkeys, called macaques, which were initially infected in captivity. Phylogenetic analysis has grouped HIV-2 in the same category as SIVmac239 since they both share a common lineage from SIVsm, which infects but does not lead to disease in the West African primate sooty mangabey (reviewed in (67)). One of the reference sequences for these HIV-2 and SIV strains is SIVmac239 (Genbank accession number M33262). HIV-1 originated from SIVcpz, which is the primate lentivirus found in chimpanzees (55); HIV-1 is further subclassified into "groups" of genetically similar isolates. These groups are determined based on geological clustering along with phylogenetic similarity and are labeled group M (main), group O (outlier), and group N (new). Group M, which is the most prevalent HIV-1 group worldwide, is further divided into subtypes A through K (reviewed in (67)). HIV-1_{NL4-3} (Genbank accession number AF324493) belongs to HIV-1 group M subtype B, and the 3' half of the genome shares a nearly identical sequence to the reference strain of that same subtype called HXB2.

Throughout the evolution of the viral sequence, HIV-1, HIV-2, and SIV isolates have maintained quite similar protein structures. Consider, for example, the lentiviral PR enzymes, whose sequences and structures are different from cellular proteases in overall length and conformation of many of their functional regions. Even between HIV-1_{NL4-3} and SIVmac239, the amino acid sequences of PR differ by approximately 50%. However, the different viral isolates maintain common structural elements, most of which can be superimposed almost identically (174). Functional regions in the viral RNA, in contrast, not only differ in sequence and length, but also in structure. An obvious example is the trans-acting response element (TAR), which forms a three-helical stem in HIV-2 and SIVmac239 but a single stem-loop in HIV-1 (12). In these and other instances throughout the viral genomes, the structure of the proteins seems to be more evolutionarily conserved than that of the RNA.

B. RNA Structure

1. Importance of RNA structure in biological systems

Ribonucleic acid (RNA), like its counterpart deoxyribonucleic acid (DNA), is a polymer chain composed of individual nucleotide monomers. Ribonucleotide monomers are linked together through a ribose-phosphate backbone with each ribose also linked to a base of either adenine (A), uracil (U), guanine (G), or cytosine (C). RNA is synthesized as a single strand of these nucleotides, as opposed to DNA which stays base-paired with its template to form a double-stranded double helix which, for animal cells, stays in the cellular nucleus. Due to the lack of a complementary strand to stabilize the RNA in solution, the nucleotides form hydrogen bonds and hydrophobic interactions to reduce

their free energy and stabilize the molecule (66). These interactions occur through Watson-Crick base pairs where U pairs with A and G pairs with C. The thermodynamic stability of each pair is determined by the number of hydrogen bonds that form between the two nucleotides, with the stronger G-C pair forming three bonds and the weaker A-U pair forming two. In a non-Watson-Crick pair, G can also pair with U by forming two hydrogen bonds. These interactions can occur when the single-stranded RNA folds back on itself to create short, irregular stretches of A-form helix. Each turn of the A-form RNA helix has 11 base pairs with each base pair rising 2.73 Å, compared to the 10 base pairs in B-form helix made by DNA (148). Variations in the pairing interactions within the RNA molecule lead to many different motifs that define the RNA structure. A single nucleotide bulge occurs when one individual nucleotide does not have a pairing partner while the surrounding nucleotides are forming a helix. A multiple nucleotide bulge has more than one unpaired nucleotide on the same side of a helix. A hairpin loop or stem-loop is a structure that forms when the RNA strand folds back in a small loop and allows nearby nucleotides to form base pairs. RNA helices form mismatch pairs when two nucleotides directly across from each other do not form canonical G-C, A-U, or G-U pairs. Internal loops are formed when these mismatch pairs include more than two nucleotides, with symmetrical loops having the same number of nucleotides on each side of the helix, and asymmetrical loops having differing numbers of nucleotides on each side of the helix. Junctions form in RNA structure when two, three, or four stems intersect. RNA also forms long-range interactions such as pseudoknots which are formed when loop regions pair with nucleotides to extend the helix of the stem; kissing hairpins are pairing interactions between two loops of separate stem-loop structures, and hairpin loop-bulge

contacts are pairing interactions between a loop from one stem and a bulge from another (reviewed in (68)).

These various pairing interactions occur, as stated above, and minimize the free energy of the system. Pairing happens spontaneously at physiological temperature and pH and is accomplished by unfavorably lowering the entropy during hydrogen bond formation but favorably and significantly lowering the enthalpy by ordering the water around the RNA molecule. This entropy loss increases as the number of consecutive base pairs increases (reviewed in (68)). The RNA folds into the most energetically favorable structure, folding into helices by forming base pairs whenever such interactions lower the free energy. The strength of these helices, however, is not determined solely by the number of G-C bonds compared to the less energetic A-U or G-U bonds. Instead, the length of the helix and the surrounding base pairs (or "nearest neighbors") play a role in determining free energy of the structure (17). Along with hydrogen bond formation, π-stacking interactions play an important role in determining the structure of nucleic acids. The stacking of the aromatic rings between neighboring nucleotide bases helps conserve the α-helical structure and increases the bases' ability to hydrogen bond with one another (116, 119). The strength of a given helix is dictated by not only base pairing interactions, but also the amount and positioning of loops, bulges, mismatches, and consecutive G-C pairs (reviewed in (68)).

RNA structure plays an important role in many biological systems. Transfer RNA (tRNA) is a classic example of the role of RNA structure in biology. Its cloverleaf structure, which further folds into two pairs of stacked helices, allows biochemical interactions to occur at both the anticodon and the amino-acid binding site, both of which

are required for precise protein synthesis to take place (reviewed in (44)). As part of larger RNA molecules, riboswitches bind to ligands or aptamers, which force structural changes and can regulate gene activity and metabolic processes (reviewed in (109)). Group 1 introns use ribozyme-catalyzed cleavage to self-splice out their own exons as is the case with certain ribosomal RNAs (rRNAs) in *Tetrahymena* (90) and can be engineered to splice a heterologous mRNA such as for human *p53*, correcting splicing defects and repairing damaged mRNA (170). The structure of ribosomal RNA allows the RNA to participate in peptide bond formation and contributes to a number of tertiary interactions which enable it to efficiently partake in protein synthesis (7, 126) reviewed in (159) and (127). RNAs that are involved in critical cell processes (and/or have catalytic function) evolve slowly and in ways that conserve the base pairs involved in the important features of secondary structure, allowing these structures to be compared over evolutionary time.

2. Structure of viral RNAs

RNA structure has been implicated in the control of various functions including transcription, splicing, aminoacylation, translation, and encapsidation in many viruses. Tobacco bushy stunt virus contains RNA stem-loop structures that modulate mRNA transcription initiation dependent on base-pairing interactions (168). A single nucleotide change in some influenza virus H5N1 strains affects the conformation of an RNA structure, shifting the balance between hairpin formation and pseudoknot formation in this region, which leads to possible alterations in splicing and impacted virulence of this strain (59). Some viruses including Nemesia ring necrosis virus and satellite tobacco

mosaic virus contain tRNA-like structures at their 3’ regions that can be aminoacylated (50, 87). Others, including members of the *Dicistroviridae* family, use these tRNA-like structures at their 5' regions to guide translation initiation as part of internal ribosome entry sites (IRESes) (31). The IRES structure of the RNA in the 5'UTR of this and members of the *Picornaviridae* family initiates translation via an internal ribosome entry site (IRES) in a (m^7Gppp) cap-independent way by competing for cellular translation factors then recruiting and binding them to an internal position in the viral RNA through their conserved RNA structural motifs (reviewed in (51, 138)). RNA structures have been shown to influence viral packaging by folding into “panhandle” hairpin structures in *Hantaviruses* (120) and by mediating dimerization and protein recognition in Moloney murine leukemia virus (33).

3. RNA structure determination: SHAPE

Determining the structure of large RNAs is challenging and has been approached by folding the RNA to obtain the lowest free energy state, comparing related RNA sequences to identify compensatory changes that can preserve base pairs, and by probing the structure with chemicals or nucleases to infer the presence of paired versus unpaired regions. Structural studies are limited by the size of the RNA. Bioinformatics-based folding programs help determine the lowest free energy of the entire structure, but do not incorporate chemical data (179). Selective 2' hydroxyl acylation analyzed by primer extension (SHAPE) is a hybrid approach able to map structure in large RNAs by using chemical analysis to constrain a folding program (118, 122, 164, 173). The 2' hydroxyl group of the ribose is much more reactive with the 1M7 reagent at single-stranded or loop

positions where the base is not paired compared to positions where the base is paired. After treatment with 1M7, the extent of derivatization at each ribose is assessed as terminations of DNA synthesis using reverse transcriptase and fluorescent primers. This approach allows determination of reactivity values for each individual nucleotide and inputs these reactivity values as pseudo-energy constraints in *RNAstructure* (146), an RNA structure prediction program. SHAPE has been used to map diverse RNA structures, including the structures near the start codons of all 13 mammalian mitochondrial RNA open reading frames (72), ribosomal RNA (40), group 1 intron RNA (45), and even other retroviruses such as Moloney murine leukemia virus (56).

4. Functions of known RNA secondary structures of lentiviruses

The RNA structure in certain regions of the lentiviral genome has been studied in great detail, and functions have been assigned to various structures throughout the genome that serve different purposes throughout the viral replication cycle. The 5' regulatory region of the genome does not code for any proteins, and is therefore termed the untranslated region (UTR). The UTR is vital to the genome because it contains many known functional structures. These include the trans-acting response (TAR) hairpin which interacts with the viral Tat protein to increase the level of transcription of full length RNA (8, 12, 18), the primer binding site (PBS) which anneals to the $tRNA^{Lys3}$, which serves as primer to initiate negative-strand synthesis during reverse transcription (85), the dimerization initiation site (DIS) which is contained in the loop region of a hairpin and makes pairing interactions with the second strand of viral RNA that is

copackaged within the virion (132, 155), and the psi stem which has been implicated in packaging of the full-length genomic RNA into the virion and in dimerization (26).

Although the coding region of the genome has selective pressure to maintain the necessary encoded amino acid sequence, the RNA downstream of the UTR still retains a sequence that is able to form critical structures. Two essential structures in viral replication are the Gag-Pro-Pol frameshift stem and the Rev-response element (RRE). The Gag-Pro-Pol frameshift stem is found in only the full-length RNA. The *gag* gene encodes a polyprotein, which is subsequently cleaved to yield the MA, CA, and NC structural proteins. Near the 3' end of the *gag* gene, a poly(U) sequence directly before the stable Gag-Pro-Pol frameshift hairpin causes the ribosome to stall and occasionally slip to the –1 reading frame. This slip shifts the reading frame of the ribosome prior to the Gag stop codon to form the Gag-Pro-Pol polyprotein, allowing the ribosome not only to translate the Gag protein (in the 0 reading frame), but also the RT, PR, and IN enzymes, which are part of the Pro-Pol region (in the -1 reading frame) (69). The downstream RRE structure is present in all incompletely spliced mRNA, including the *vif*, *vpr*, and *vpu*/*env* message, and full-length genomic RNA containing *gag* and *gag-pro-pol* genes. The RRE secondary structure allows for transport of these transcripts out of the nucleus via recognition by the viral Rev protein (47, 49, 105, 106) in a Crm1-dependent pathway in which the host nuclear export factor Crm1 helps to transport the mRNA through the nuclear pore complex into the cytoplasm (32, 52, 139).

The functional RNA structures contained within the TAR, RRE, and frameshift stem regions fold in ways that facilitate RNA-protein interactions. TAR binds to the viral protein Tat via a U-rich bulge that interrupts a stem-loop structure. The TAR structure

undergoes a rearrangement when the arginine side-chains of Tat are bound, creating a binding pocket for the protein and allowing contact to the other basic residues of Tat (reviewed in (79)). The viral Rev proteins cooperatively bind to the helical structures that compose the RRE. The initial binding interaction occurs at stem-loop IIB of the RRE through binding of a single Rev monomer (27, 34, 162). Subsequently, more Rev proteins cooperatively assemble via hydrophobic interactions between Rev molecules and electrostatic interactions between the proteins and the RNA (35, 38, 84, 110). Unlike the TAR and RRE interactions with viral accessory proteins, the RNA structure at the Gag-Pro-Pol frameshift stem interacts with cellular factors that compose the ribosome. Instead of binding sites that encourage interaction, the function of this stable stem is to hinder translation by the ribosome, discouraging macromolecular interactions between the RNA structure and the translation machinery (69). The functions of RNA structure described above are performed by only a fraction of the total lentiviral RNA. Analyzing other conserved RNA structures may divulge important roles and interactions that are yet unknown throughout lentiviral replication.

C. Pre-mRNA splicing

1. General principles of splicing

The initial RNA transcript can be modified in several ways including alteration through removal of long segments of the RNA by splicing. Splicing occurs in the nucleus and is executed by the spliceosome, a complex machine composed of both protein and RNA (reviewed in (166)). This mechanism removes segments of RNA, called introns, and combines the remaining exons to form messenger RNA (mRNA), the mature form of

the RNA, to be used as a template for protein translation. In total, about 95% of human cellular RNAs are spliced, and the genes have an average of seven exons (135). Each intron starts with a 5' splice site (5'ss) beginning with the sequence GURAGU to serve as the donor, and a 3' splice site (3'ss) ending with the sequence YAG, which is preceded by a polypyrimidine tract (py-tract), to accept the ligation event from the donor (Figure 1.2a). The RNA is cut at the donor site and the 5' end of the intron is transferred to the 2'hydroxyl of a conserved adenosine at the branch-point sequence YNYURAY (Figure 1.2a) by a transesterification reaction. The 3'ss is then broken after the AG and linked to the upstream exon through another transesterification reaction (102, 143). The spliceosome uses these common sequences to aide in recognition during the splicing events. Different small nuclear RNAs (snRNAs) combine with proteins to form small nuclear ribonucleoprotein (snRNP) complexes. These complexes recognize and bind to the splicing donor, acceptor, or branch-point sequences sequentially. The U1 snRNP recognizes the 5'ss while the U2 activating factor (U2AF) binds the 3'ss. These elements then act cooperatively to recruit U2 snRNP (Figure 1.2a). Although these factors work to excise the introns, these assembling complexes do not recognize the introns in their entirety. Instead, they bind to and define both sides of the individual exons before joining them together (11, 57, 147).

When a single transcript is spliced in two or more alternative ways, it can give rise to more than one protein or protein isoforms with different activities. This alternative splicing is accomplished in different ways: *i*) The same transcript is spliced at completely different exons leading to completely different proteins, *ii*) the spliceosome intermittently recognizes and includes an exon in the middle of the transcript, adding inserts to the

middle of a protein, or *iii*) multiple exons are available and the machinery can chose any of them to include in the message (115). Exons range in size and can potentially be rather small. When the gene includes even a few extra residues due to alternative exon inclusion, the protein structure and function can be drastically affected. In this way, the approximately 25,000 genes in the human genome can give rise to around 500,000 proteins, or a short viral genome can generate the many necessary viral proteins for efficient replication, allowing greater diversity from a small amount of genetic information.

2. Regulation of splicing

The positioning of introns relative to the genetic code appears to be random. However, the ability of the spliceosome to recognize these introns and splice them properly is vital to proper cellular function. The spliceosomal machinery, especially the snRNPs, must be able to distinguish genuine splice sites from cryptic or fortuitous sites. The RNA binding factors must also be able to recognize and use the correct alternative sites given the needs of the cell or virus. Therefore, many factors are involved to assure this process is precise. The spliceosome accurately splices the transcribed RNA based on many factors that are encoded in the mRNA itself. These are displayed as sequences termed splicing regulatory elements (SREs) and include exonic splicing enhancer (ESE) sequences which enhance splicing of the exon they are part of, exonic splicing silencer (ESS) sequences which are located in the exon but inhibit splicing at a certain acceptor or donor site, intronic splicing enhancer (ISE) sequences which amplify a splicing event that eventually eliminates their sequence from the mature message, and intronic splicing

silencer (ISS) sequences which are also located in the potentially discarded intron but act to inhibit such a splicing event from happening (9, 53, 163).

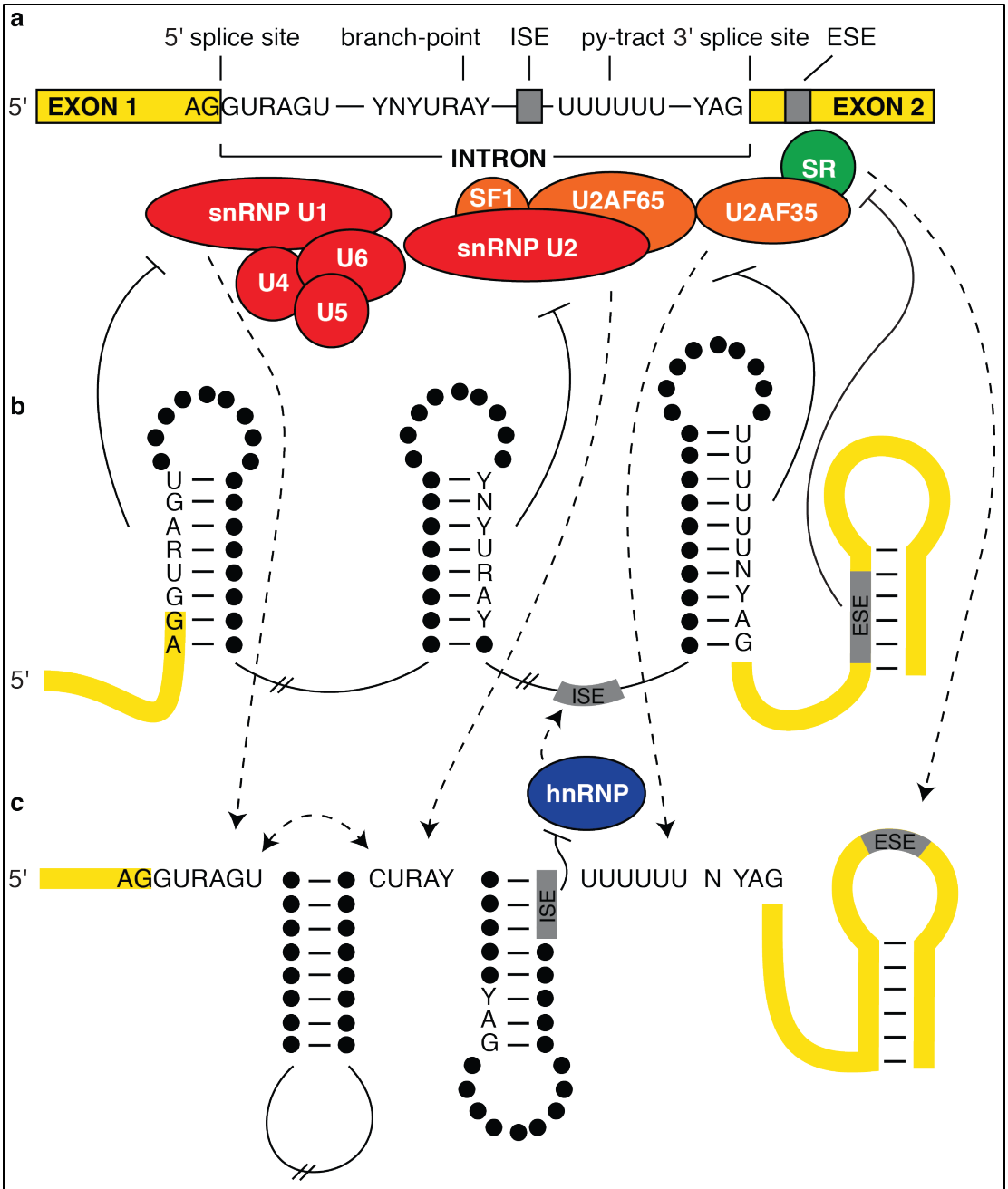


Figure 1.2. Splicing factors and the impact of RNA structure on splicing regulation. Figure adapted from (169). (a) Generic splicing factor recognition sites in exons (yellow) and introns. Letters represent nucleotides as described in text. Pyrimidines (Y), purines (R), and any nucleotide (N) are indicated. Splicing factors are indicated as ovals at interaction sites of snRNPs (red), U2AFs (orange), SR proteins (green), and hnRNPs (blue). (b) RNA stem structures are shown secluding sequences from protein recognition. (c) Recognition sequences in single-stranded and loop regions enhance protein binding. YAG indicates a cryptic splice site, which is secluded in a hairpin. Hashed loop indicates a large structure, which is paired at the base to bring recognition sequences into proximity.

Functional SREs seem to inhabit single-stranded regions more often than base-paired regions, allowing binding proteins to identify and interact with these particular sequences. It has been found that enhancer-dependent regions (those with weak splicing sequences) are particularly marked with single-stranded ESE motifs, and regions that depend on silencing have a stronger occurrence of single-stranded ESS motifs (65). These SRE sequences work by binding specific proteins, which in turn inhibit or enhance splicing factors from recognizing the 5'ss or 3'ss sequences in close proximity. The acceptor sites are enhanced by the ability of surrounding sequences to bind SR proteins (Figure 1.2a), which are defined by their RNA recognition motifs and their carboxy-terminal domain that contains several serine and arginine repeats (reviewed in (58, 156). These proteins interact with different parts of the spliceosome: regions of the U1 snRNP, U2AF, and U4/U6.U5 tri-snRNP (reviewed in (58)). Interactions between parts of the spliceosome and the corresponding splice sites are inhibited by ISS or ESS sequence recognition by heterogeneous nuclear ribonucleoprotein (hnRNP) complexes (Figure 1.2b), blocking recognition and disallowing splicing at a given site (41). In the cell, however, the splicing factors, including SR proteins and hnRNP complexes, are present at varying concentrations, which leads to the concept that cells are able to regulate the utilization of different splicing pathways by controlling the amounts of these complexes. This allows for more complexity in splicing regulation through cellular control factors that dictate the ratios of splicing enhancement and silencing factors (156).

RNA structure itself can have a regulatory effect on splicing. In a broad sense, structures that form within large introns to bring distant splice sites closer together allow the assembling spliceosome to recognize all of the necessary sequences within a shorter

range (24) (Figure 1.2c). At a closer scale, structure formation occurs co-transcriptionally, allowing the RNA to make short-range interactions as each nucleotide is added to the polymer (150). These short interactions, particularly stem-loop structures, have been implicated in controlling recognition of splicing factors (Figure 1.2b and c). The minimal structure needed is a small stem-loop as short as 7-bp, which is adequate to enclose and seclude an enhancer sequence, impeding its activity (98). An analysis of splicing in mammalian cells showed conserved structures around splice sites that conceal certain sites where regulation is necessary for control of gene expression and repression of various disease phenotypes (152). Furthermore, GC content, which is implicated in stronger pairing interactions, has been shown to be enriched around alternative splice sites, particularly those for the first possible exon, suppressing usage of such sites by the splicing machinery (178). These analyses imply that stable secondary structures around splice sites serve to seclude these sequences from their binding proteins and allow other alternative sites to be used at higher frequency. A stem that includes the py-tract and the AG 3'ss precludes binding of that sequence to U2AF, while a stem that makes pairing interactions with the sequences involved in the 5'ss precludes U1 snRNP binding (Figure 1.2b). Even though the U2AF and U1 snRNP help recruit U2 snRNP, if the branch point sequence is paired in a stem structure, the binding of U2 snRNP to that site will be disrupted. In contrast, when these sequences are in single-stranded or loop regions, they increase the binding interactions to their respective proteins (Figure 1.2c). Additionally, cryptic acceptor sites, when paired in structure, are hidden from U2AF recognition and binding (reviewed in (169)) (Figure 1.2c).

Splicing regulation depends not only on the accessibility of actual splice sites but also the RNA structures around *cis*-acting SRE sequences. These motifs bind certain proteins that allow or disallow splicing to occur. ESE and ISE motifs bind to SR proteins which, when bound, enhance the splicing efficiency of the nearby 3'ss. If enhancer elements are secluded in base-pairing interactions, the SR proteins are not able to bind and thus do not impact splicing at the 3'ss (reviewed in (169)) (Figure 1.2c). In a converse event, an hnRNP can bind to either an ISS or ESS (Figure 1.2b). Once bound, this causes the structure around the exon to form a loop and disallows U1 snRNP binding at the 3'ss near the silencer motif, thus avoiding the given exon altogether (125), (reviewed in (19)). Although they do not incorporate the given regulatory element, structures that occur upstream of SRE regions have been shown by computational analysis to enhance the function of the given SRE. These stable structures near the region of regulation could potentially function to interfere with any RNA conformation that might have otherwise been part of the SRE (98). Taken together, these concepts strengthen the idea that structure of the pre-mRNA around the splice site sequences, which includes enhancer and silencer elements, helps determine whether the given site will be recognized and spliced by the spliceosome (reviewed in (169)). The general model for splicing regulation by RNA structure is the following: The single-stranded SREs are recognized by and bind their regulatory proteins, and the base-pairing of these sequences limits their effectiveness.

3. Pre-mRNA splicing of HIV-1

Cellular genomes are vast in size compared to viral genomes, especially those of RNA viruses. These RNAs need to strike a balance between being small enough to allow efficient replication in a cell yet large enough to contain all of the genetic material to produce the necessary viral proteins. One way lentiviruses accomplish this is by their ability to be spliced in a multifaceted manner. The proviral DNA is transcribed by cellular polymerase Pol II to produce a long unspliced RNA that can either be used as the genomic RNA in the packaged virus, as a transcript for Gag and Gag-Pro-Pol translation, or spliced to produce the over 40 different mRNA species used to generate the remaining viral proteins (142). For HIV-1, the cellular splicing machinery utilizes four 5'ss donors and eight 3'ss acceptors (Figure 1.3a).

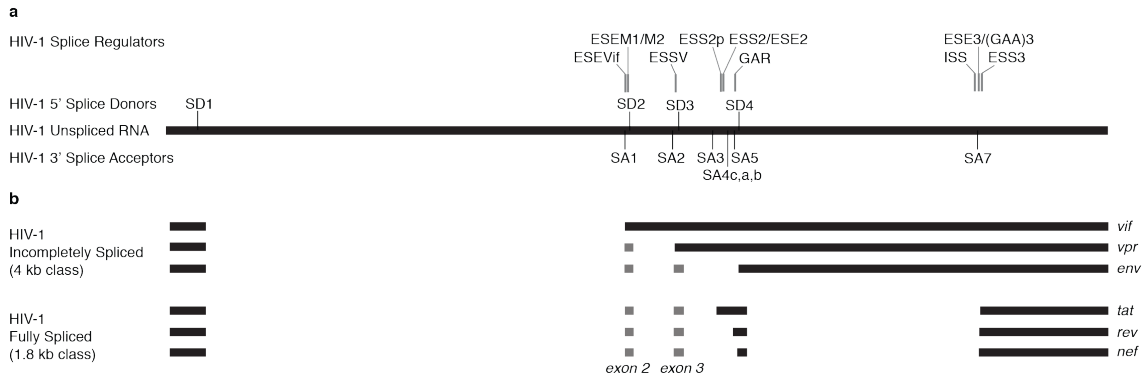


Figure 1.3: Schematic showing splice sites and regulatory regions in HIV-1. (a) Location of splice donors SD1-4 and acceptors SA1-7 in relation to unspliced RNA. Known splice regulatory regions are shown above. (b) Exons producing the incompletely spliced (4 kb class) and fully spliced (1.8 kb class) mRNAs. Exons 2 and 3, which are sometimes included in the mature transcripts, are gray.

The first 5'ss SD1 is the major splice donor and is used in all spliced products. When SD1 is used in conjunction with the splice acceptors SA1, SA2, or SA5 with no downstream splicing events, the *vif*, *vpr*, or *vpu/env* mRNA transcripts are produced, respectively. These are categorized as the 4 kb class of transcripts (or singly spliced mRNAs) because they incorporate the RRE-containing sequence between SD4 and SA7 (Figure 1.3b). The splice variants that excise this intron are grouped as the 1.8 kb class of transcripts (or small, multiply spliced mRNAs), which is comprised of mRNAs spliced once from SD1 to SA3, SA4(a,b,c), and SA5 and then again from SD4 to SA7 creating *tat*, *rev*, and *nef* mRNAs, respectively (142) (Figure 1.3b). Although the transcripts that are spliced directly from the SD1 donor to the closest upstream acceptors are the most abundant variants, smaller exons can still be incorporated into the mRNA, including exon 2 from SA1 to SD2 and exon 3 from SA2 to SD3 (Figure 1.3b). The start codons for Vif and Vpr are after SD2 and SD3, respectively. Therefore, any splicing event that uses these donors effectively excises those start codons, producing the corresponding transcript of the next acceptor that is used (142).

An important feature of unspliced and singly spliced (4 kb class) mRNA is the RNA structure in the SD4-SA7 intron, the RRE. Typically, unspliced or incompletely spliced mRNA remains in the nucleus until it is fully spliced (32). If incompletely spliced pre-mRNA interacts with U1snRNA alone with no other splicing factors, it is retained in the nucleus and disallowed to either be fully spliced or exit into the cytoplasm (21, 95). HIV-1 overcomes this nuclear retention by encoding the nuclear-export protein Rev. Once Rev is translated in the cytoplasm (using the multiply spliced and efficiently transported small mRNA), the Rev protein then returns to the nucleus, recognizes and

binds the RRE of these incompletely spliced mRNA molecules, and shuttles them out of the nucleus (32, 139).

4. HIV-1 splicing regulation

The diverse mRNA splicing that occurs in HIV-1 transcripts is a result of complicated alternative splicing (142). The splicing events must be highly regulated to allow the necessary amounts of each transcript to be produced. For example, if the nascent transcript were completely spliced to the smallest mRNA form, or if any of the splice sites were preferentially used, one protein product would dominate. Instead, the pre-mRNA contains many weak splice acceptor and donor sites along with weak branch-point sequences, leveling the potential for all to be recognized by the splicing machinery. The acceptor sites contain purines, causing them to be weaker than the ideal polypyrimidine tracts that enhance splicing (4, 46, 129, 153, 157). Maintaining the balance for splicing at each acceptor remains complicated, however, because the RNA also contains *cis*-acting regulatory elements in the form of sequences and secondary structures that enhance or repress splicing at particular sites (reviewed in (160).

Regulation of HIV-1 RNA splicing begins with the major splice donor (SD1). This donor has the capacity to splice to any of the downstream acceptors, and when splicing occurs within the primary transcript, the splicing machinery will always use SD1 (142). The SD1 sequence is composed of the loop region of a conserved stem-loop structure in the 5' region of the genome which, when mutated, alters the efficiency of splicing using SD1 (2). SD1 becomes a less efficient 5'ss when the structure around SD1 is changed to incorporate the SD1 recognition sequence into a hairpin structure,

suggesting that accessibility of SD1 in a loop is imperative for SD1 recognition (2). The mechanism that seems to be at work in this region is one that allows the SD1 site to be visible to the splicing machinery by placing it in a loop instead of in a paired interaction.

Splicing to the first splice acceptor SA1 with no further downstream splicing events results in the mRNA transcript for the Vif protein. This protein is necessary for cellular APOBEC3-G and -F cytadine deaminase downregulation, therefore diminishing the cellular restriction factors' mutagenic effects on the newly reverse-transcribed negative strand DNA (29). Regulation of the SA1 site is crucial for usage of downstream splice acceptor sites. This site has been determined to include the strongest splice acceptor site of all the HIV-1 3'ss sequences (76). However, this 3'ss must be used in moderation to allow for splicing from SD1 to the other possible splice acceptors.

Following the idea that splicing complexes recognize the entire exons instead of the introns (11), it has been shown that production of the *vif* message is regulated at the second splice donor site (SD2) (48, 76), located 50 nucleotides downstream of SA1 and 78 nucleotides upstream of the *vif* start codon. If used, SD2 splices to a downstream 3'ss, creating a small exon 2 and excluding as an intron the sequence that would begin Vif translation (142). In this way, even if SA1 is the acceptor used for SD1, the *vif* message is not necessarily the final product. A weak 5'ss sequence at the SD2 site has been shown to enhance the production of the *vif* message (104, 108). Another enhancer, ESEVif, occurs in the region between SA1 and SD2 along with a GGGG splicing silencer that follows SD2. These are in competition with one another to systematically regulate splicing at this site and, ultimately, expression of Vif (48). The regulatory sequences ESEM1 and ESEM2 have also been described and perform similar functions to ESEVif in enhancing

the use of SD2 (76). These elements have been identified through mutagenesis analysis that specifically disrupted the sequences, however, this was done with no regard to the structures that may have been formed in these regions. As of yet, the RNA structure around SA1 has not been implicated in any silencing or enhancing function for splicing at this region.

A splicing event to the second splice acceptor site (SA2) produces a message that encodes Vpr, a protein important for both promoting infection in myeloid cells (6, 28) and arresting the cell cycle of dividing cells (64, 74). A regulatory sequence termed ESSV has been identified in this region that helps to repress the use of SA2 (13). This silencing mechanism functions when the ESSV sequence binds hnRNP A/B proteins and inhibits binding of U2AF65 to SA2 (42). Specific mutagenesis of ESSV has localized the interacting element to a 16-nt sequence (103).

The splicing events that use SA3 will produce the *tat* transcript. The usage of SA3 is usually followed by excision of the intron from SD4 to SA7, and this is a requirement for the production of functional Tat (142). The Tat protein recognizes the TAR hairpin structure in the 5' UTR and recruits transcription factors to upregulate RNA Pol II function during mRNA transcription (15, 93). However, Tat has an apoptotic effect on the cell (16, 79, 97) and its production is therefore well regulated. An upstream silencer called ESS2p (70) and ESS2, a silencer downstream of SA3 (153), work against a relatively stronger py-tract compared to most others on the RNA (71) and splicing enhancer sequence ESE2 (176) to accomplish this regulation. These sequences are located on two stem-loop structures SLS2 (which contains ESS2p (70)) and SLS3 (which contains both ESS2 and ESE2 located directly next to each other (3, 176)), which reside

in the vicinity of SA3. Various regions on these stem-loop structures interact with enhancing and silencing proteins to modulate splicing at SA3. The proteins recognize sequences within these RNA structures that seem to be shared binding sites between many of them, thus leading to competition of regulation at this site (61).

Splicing from SD1 to any of the splice acceptors SA4c, SA4a, and SA4b followed by the SD4-SA7 splicing event leads to the *rev* mRNA transcript. Use of SA5 by SD1 leads to the transcript for either *nef* or *vpu*/*env* depending on if the intron between SD4 and SA7 is excised (*nef*) or included (*vpu*/*env*) (142). An ESE element termed GAR is located directly downstream of SA5 and has been shown to regulate usage of the splice acceptor sites SA4c/a/b and SA5 and the downstream donor SD4 (20, 75). This regulation is accomplished in two ways: binding SR proteins in a bidirectional manner to allow sufficient usage of the upstream acceptors, and binding U1 snRNP to SD4 which amplifies expression of unspliced and incompletely spliced mRNA (75). An RNA structure at this site has yet to be identified.

Perhaps one of the most intuitive splicing regulatory regions is that pertaining to the splicing event that occurs between SD4 and SA7. When this intron is excised, the 1.8 kb class of mRNA products is produced. Even in the absence of Rev, these products are able to exit the nucleus freely. If the SD4-SA7 intron is not spliced from the transcript, the 4 kb class of mRNAs and the unspliced product are generated and require Rev for efficient transport from the nucleus into the cytoplasm. The sequences that serve as SREs around this splicing interaction have been studied in detail, and the RNA structures that include these SREs are known. An ISS at this site consists of a sequence that makes up one half of the hairpin structure SLS1, the two-part ESS3 (ESS3a/b) sequence is base-

paired in the structure SLS3, and ESE3/(GAA)3 has been described as a large bulge region in the structure SLS2 (37, 111). As with the different silencers and enhancers that act upon SA3, these various SREs must cooperatively and competitively bind different hnRNP factors and SR proteins to regulate splicing at this site (111).

D. Thesis Overview

Regions of RNA secondary structure play essential roles in the replication cycle of HIV-1. The SHAPE technique has been applied to determine the RNA secondary structure of the full-length HIV-1_{NL4-3} genome, and this analysis has shown many elements of RNA structure (172), but only a fraction of these have been previously studied. One tool to assess the importance of these structures is to determine the extent to which they are conserved over evolutionary time and the extent to which they are maintained after mRNA splicing.

The second chapter describes the application of SHAPE technology to develop a secondary structure model for the genomic RNA of a second primate lentivirus, simian immunodeficiency virus (SIVmac239), which shares 50% sequence identity at the nucleotide level with HIV-1. In both genomes approximately 60% of the nucleotides are paired within the coding region (8,738 nucleotides). However, only about half of these paired nucleotides are paired in both sequences, and only 58 base pairs form with the same pairing partner in the coding region of both sequences. Thus on average the RNA secondary structure is evolving at a much faster rate than the sequence. Some structures are conserved between HIV-1 and SIVmac239, including in the 5' untranslated region (5' UTR), the Rev responsive element (RRE), a pseudoknot to sequester the 5'

polyadenylation sequence, the polypurine tracts (PPT and cPPT) that begin plus-strand synthesis, and the stem-loop structure that includes the first splice acceptor site. Structure at the Gag-Pro-Pol frameshift site is maintained but in a significantly altered form. As with all lentiviruses, the HIV-1 and SIVmac239 genomes are adenosine-rich and cytidine-poor. Approximately two-thirds of the cytidines, uridines, and guanosines are base-paired while only one-third of adenosines are base-paired, leading to the concentration of adenosines in single-stranded regions (55% of the unpaired nucleotides). Thus the base composition of the structured regions is very different from either the unpaired regions or the genome as a whole. Structures with adenosine content equal to or greater than the number of guanosines had higher SHAPE reactivity and were not conserved between the two genomes. By contrast, those structures in which guanosines were more abundant than adenosines had lower SHAPE reactivity and structure was maintained, although still undergoing significant evolution. This leads to the conclusion that much of the secondary structure reflects pairing in a state which allows the RNA to form and reform interactions throughout evolution of the sequence. However, regions of the structure that perform necessary functions within the viral replication cycle seem to have a high guanosine content, which stabilizes these structures and allows them to remain intact even through the course of sequence evolution.

The work in the third chapter examines regulation of splicing due to RNA secondary structure in the HIV-1_{NL4-3} transcript mRNA. I evaluate the importance of an evolutionarily conserved stem-loop structure whose pairing interactions at the base of the stem were kept constant between HIV-1_{NL4-3} and SIVmac239 genomic RNA structures. This stem-loop structure is at the first splice acceptor site SA1, and is termed SLSA1.

Mutations to this stem that disrupted the pairing interaction while keeping surrounding ESE sequences intact as well as the corresponding amino acid sequence were introduced to the HIV-1_{NL4-3} genome, creating the mutant virus SLSA1m. In a virus coculture assay, the wild-type virus outcompeted the SLSA1m by a small margin. Separately, the mutant and wild-type viruses were passaged in cells and the mRNA profiles from these cells showed a difference in splicing pattern. Taken together, these data indicate a decreased viral fitness to SLSA1m and a change in usage of the splice sites based on the disruption of this stem structure. To examine other splicing regulatory features in the context of entire transcripts of fully spliced and partially spliced mRNA, I performed SHAPE analysis on *in vitro* transcribed RNAs representing the most abundant versions of spliced mRNA for all of the viral proteins. These structures exhibit maintenance of known motifs around splice sites SD1, SA2, SA3, and SA7, but with slightly altered conformations, emphasizing the importance of analyzing these structures and pairing interactions in a whole-molecule context. I observed maintenance of some previously unreported structures around the known SRE sequence at SA4c/a/b, SA5, and SD4, implying a role of RNA structure in regulation of splicing at this region. Many of these known and newly identified structures are preserved even after splicing events excise large regions of sequence, however, some structures are altered based on initial splicing events. This leads to the conclusion that most RNA regulatory structures affecting splicing of HIV-1 are formed through local interactions and are thus made impervious to large sequence changes or deletions because of the need to maintain these structures intact in the mRNA after the initial splicing event, but some structures are altered to modify the occurrence of downstream splicing events.

In the fourth chapter, I will summarize the results of my thesis work and discuss areas where further research is needed.

CHAPTER 2

COMPARISON OF SIV AND HIV-1 GENOMIC RNA STRUCTURES REVEALS IMPACT OF SEQUENCE EVOLUTION ON CONSERVED AND NON-CONSERVED STRUCTURAL MOTIF[1]

A. Introduction

RNA secondary structures play fundamental roles in the replication of all positive-strand RNA viruses. Because of their small genomes (which are largely devoted to encoding viral proteins), these viruses use available sequence space highly efficiently. The genomic RNA of viruses forms structures necessary for various functions. For example, internal ribosome entry site elements interact with the cellular translation initiation machinery, diverse structural signals direct packaging of viral RNA into viral particles, and RNA structure can provide control signals for differential viral gene expression. The human immunodeficiency virus type 1 (HIV-1) is no exception and well-characterized RNA structures within the coding domains of the genome play critical roles in regulation of replication. These include a structure in the *env* gene, the Rev response element (RRE), that binds the viral protein Rev leading to the transport of unspliced and singly-spliced viral mRNA out of the nucleus (80, 130), and a hairpin structure preceded by a poly(U) slippery sequence that mediates a frameshift during synthesis of the Gag-Pro-Pol polyprotein (136). The untranslated regions (UTRs) of HIV-1 and simian immunodeficiency virus (SIV) contain the TAR hairpin, which recruits the Tat protein to

[1] Kristen Dang and Christina Burch contributed the data that is shown in Figure 2.2c.

modulate transcription (63, 124) (reviewed in (80)) and other stem-loop structures that are important for dimer initiation (DIS) (155), splicing (123, 142), and viral RNA packaging (10, 62) (reviewed in (101)). Several lines of evidence emphasize that the HIV-1 genome contains extensive RNA secondary structures whose functional roles are not yet fully understood (99, 167, 172).

The structures of large RNAs, like viral RNA genomes, are too complex to be predicted with confidence from first principles or thermodynamic-based algorithms alone. Useful working models can often be obtained when additional information is used to restrain the number of possible secondary structure elements. Two such approaches are to compare evolutionarily related sequences to identify RNA motifs that co-vary to preserve base pairs, and to experimentally probe the RNA structure with chemicals or nucleases to infer the presence of paired versus unpaired regions. In the selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) chemical probing approach, nucleotide reactivities show a strong inverse correlation with the probability that a nucleotide is base-paired. SHAPE-directed prediction of RNA folding has been used to develop secondary structure models for diverse RNAs (40, 45, 56, 72, 173) including the full-length genomic RNA structure of HIV-1_{NL4-3} (172). This HIV-1 model shows a very strong correlation between regions that can be targeted by siRNAs to inhibit viral replication (99) and regions that are predicted to be single-stranded, suggesting that global structural features are likely correct.

One approach to evaluating the broader significance of these structures is to examine the conservation of these structures in a related virus. To this end, we analyzed the secondary structure of the genomic RNA of a second primate lentivirus, SIVmac239, a

representative of the SIVsm/HIV-2 lineage of primate lentiviruses. HIV-2 evolved from a different primate reservoir than did HIV-1. HIV-2 arose in the sooty mangabey (*Cercocebus atys*), and SIVsm has also infected rhesus macaques in primate centers and to cause an AIDS-like illness. SIVmac239 (144) now serves as a prototype reference sequence for comparative analysis of the HIV-2/SIVsm lineage (22). SIVmac239 has a large evolutionary distance from HIV-1, and conservation of structures between HIV-1 and SIVmac239 represents an especially stringent test for functional relevance. In this analysis, we describe areas where RNA structure is maintained between HIV-1_{NL4-3} and SIVmac239, where it is divergent, and outline possible mechanisms for understanding the interplay between rapid sequence evolution in the context of selection for maintenance of function of RNA structural motifs.

B. Results and Discussion

1. Features of the SIVmac239 RNA structure

To develop an experimentally-based secondary structure model for the genomic RNA structure of SIVmac239 (GenBank accession M33262), we used a strategy similar to that used to develop a model for the secondary structure of genomic HIV-1 RNA (172). Viral RNA was purified from SIVmac239 particles and derivatized with the SHAPE reagent 1-methyl-7-nitroisatoic anhydride (1M7) under physiologically relevant conditions to discriminate between single-stranded (generally reactive) positions versus (unreactive) nucleotides constrained by base-pairing or other interactions (118, 122, 164, 173). The derivatized positions were identified as terminations of DNA synthesis by reverse transcriptase (primers listed in Table 2.1). SHAPE reactivities were measured for

9,605 nucleotides, 99.6% of the genome. These data were used as pseudo-free energy change constraints to direct RNA secondary structure prediction. In the secondary structure model for the SIVmac239 RNA genome (Figures 2.1 and 2.2), 4,970 nucleotides were predicted to be base-paired (51.5%), whereas 4,676 nucleotides were predicted to be single-stranded (48.5%).

Name	Primer Sequence
SIV309	TCCTTCAAGTCCCTGTTCAGGC
SIV443	AACCGGAGGCCTCTTCCTCTCC
SIV593	CTTTCCGTTGGGTCGTAGCCT
SIV897	GATGGTGCTGTTGGTCTACTTG
SIV1193	GTCCTTGTTGTGGAGCTGGTTG
SIV1475	GTTTGAGTCATCCAATTCTTTAC
SIV1761	TCCAGCATCCCTGTCTTCTTG
SIV2095	CTGTATCCAGTAATACTTCTAC
SIV2138	GTGGACCTAACTCTATTCCTG
SIV2386	GCCACTGCTTCAATTTTGGTCC
SIV2674	CTAGAGGTATGGAGAAATATGC
SIV2952	CCCTATGCTATTCAAGAGTTCC
SIV3225	CTCATATTCTGCTTCTGCCATC
SIV3505	CCCATACATCCTTCTCAACTGG
SIV3780	CCCTGAGTCTGTCAATGCCATG
SIV4027	CATGTTCTTCTTGTGCTGGCTC
SIV4289	AATAGTGCTGTCTGTCTTCCTG
SIV4576	GAGTCATATCCCCTATTCCTCC
SIV4831	AACTGCTATCCACCTCTTTTCC
SIV5112	TAGTTTGGTGTTACATCTGTCC
SIV5382	CCGCCTCTCTGTTTATCTCCTC
SIV5647	TGTGGTCCTTCATTTTCTGGAG
SIV5904	TAGAGGGCGGTATAGCTGAGAG
SIV6184	ATTGTCGCATTCCTCCAAGCTG
SIV6350	CTCAAAGAGTTGCCATACATCC
SIV6635	TGCAGATGACCAAGTTTCATTG
SIV6894	AGCCAAACCAAGTAGAAGTCTG
SIV7170	CAGTATACCTGGGATGTTTGAC
SIV7462	AGACTGGTCACTGTGGAGTTAC
SIV7745	CCCAGCCAATAAAGTTCGGGAC
SIV8001	AGTCAACCTTTCGCTCCCACTC
SIV8261	GAAATAAGAGGGTGGGGAAGAG
SIV8536	TGTAGGTAGGTCAGTTCAGTCC
SIV8830	CCAAGTCATCATCTTACTCATC
SIV9107	TCATCCTCCTGTGCCTCATCTG
SIV9282	TAGCCTTCTTCTAACCTCTTCC
SIV9485	GAACCTCCCAGGGCTCAATCTG
SIV9621	TTTTTACTTCTAAAATGGCAGC

Table 2.1: Sequences of primers used for SHAPE analysis of HIV-1 mRNA. Numbers indicate the 5' position in the SIVmac239 genome to which each anneals.

Figure 2.1 (Next Page): Model for the structure of the SIVmac239 RNA genome as determined by SHAPE probing and directed RNA structure refinement. The genome is divided into (a) 5' and (b) 3' halves. Colors of nucleotides indicate SHAPE reactivity on the scale shown on the left. Each sphere corresponds to a nucleotide, and side-by-side spheres indicate a base pair. Protein coding region boundaries are indicated by letters with the code shown at the bottom. Splice acceptor and donor sites(165) are labeled SA and SD, respectively. $tRNA^{Lys3}$ interaction is shown in gray. Heavy blue bars indicate base pairs in stems that are conserved between codon-aligned SIVmac239 and HIV-1_{NL4-3} RNA structures (71 total pairs). Areas of structure with a median reactivity below 0.3 over a 75 nucleotide window are numbered in green and correspond to motif numbers in Figure 2.3b. All positions are numbered in reference to the GenBank accession number M33262 for SIVmac239. A full structure, including nucleotide identity, is shown in Figure 2.2.

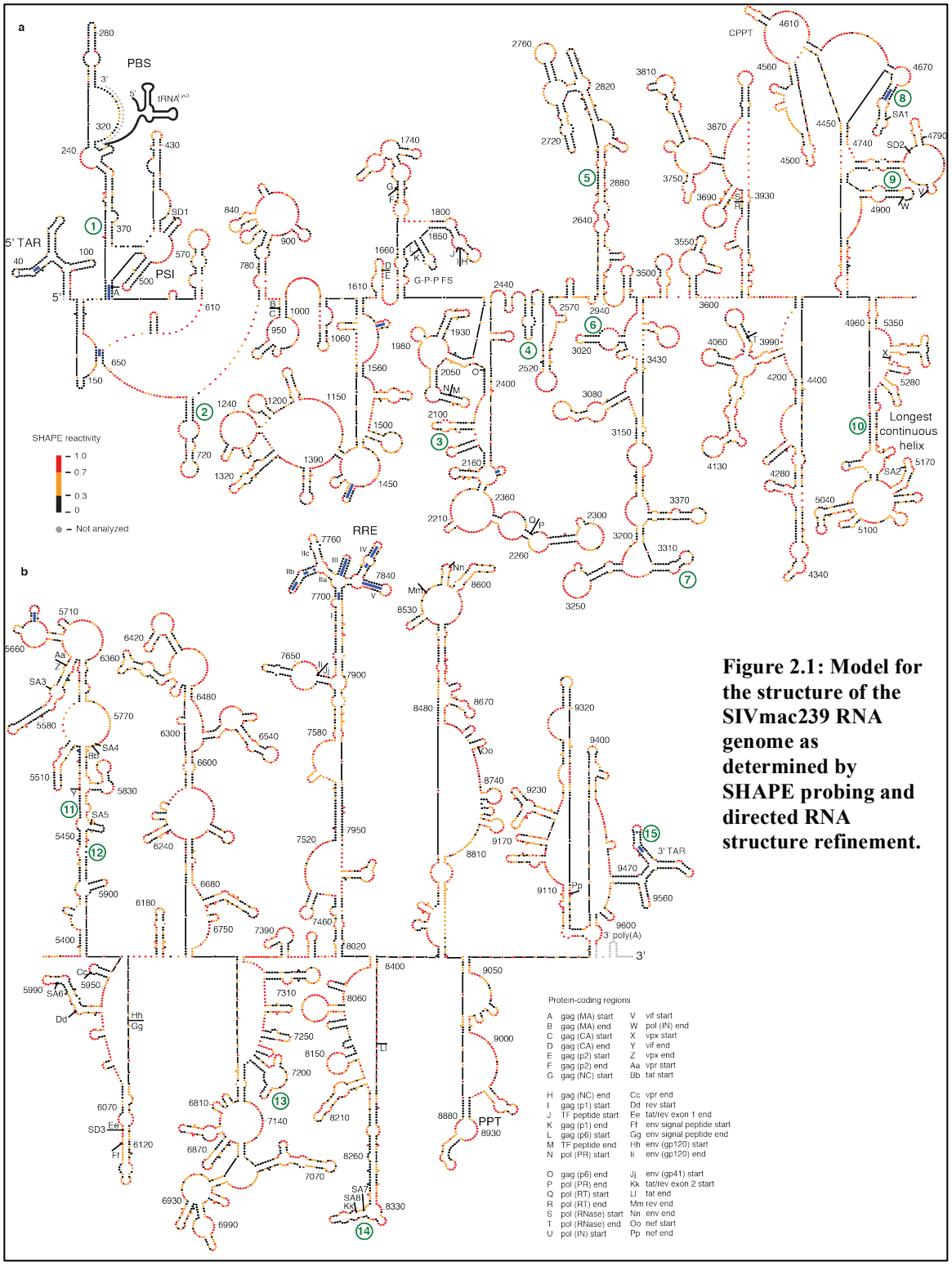


Figure 2.1: Model for the structure of the SIVmac239 RNA genome as determined by SHAPE probing and directed RNA structure refinement.

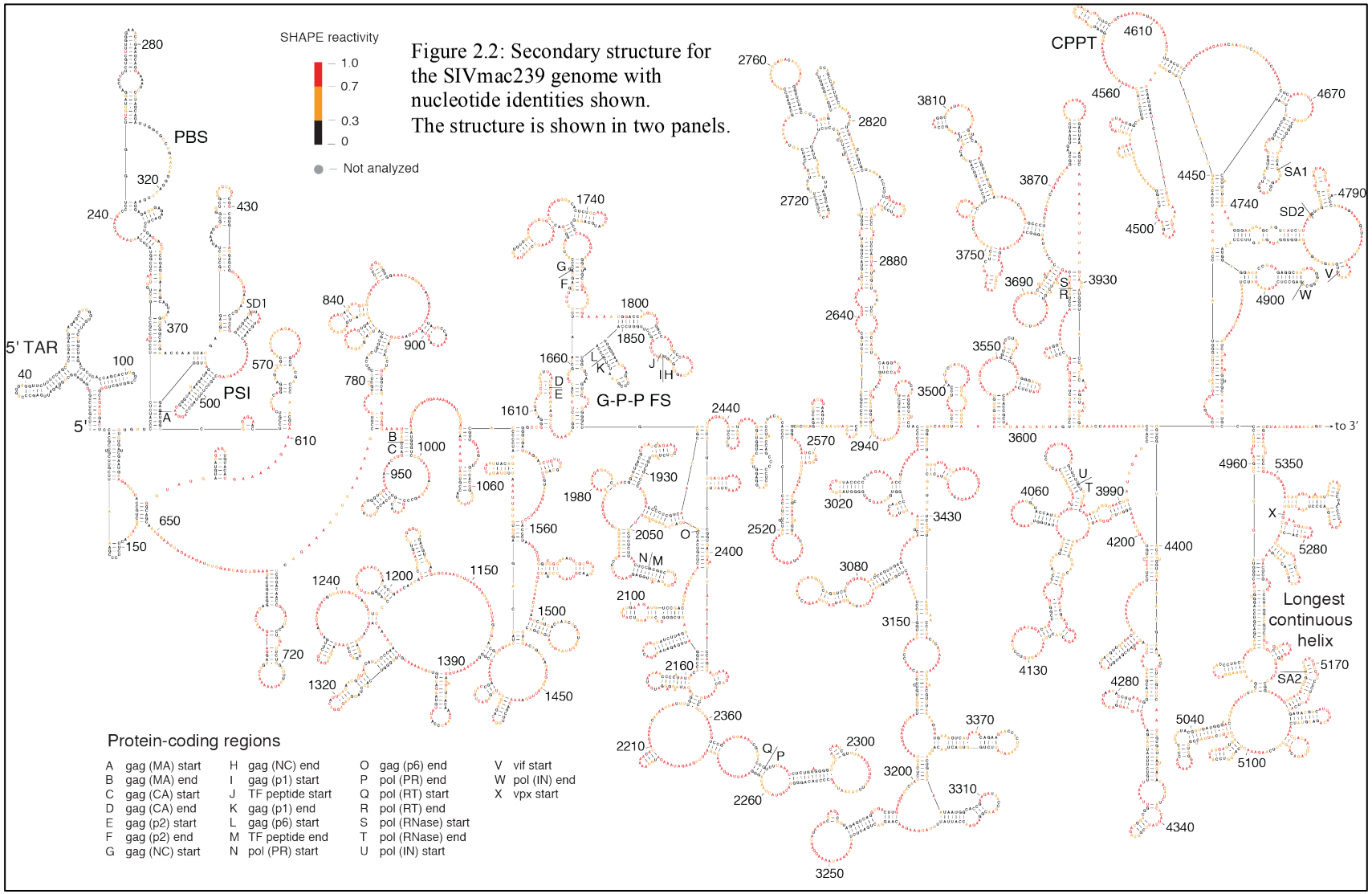
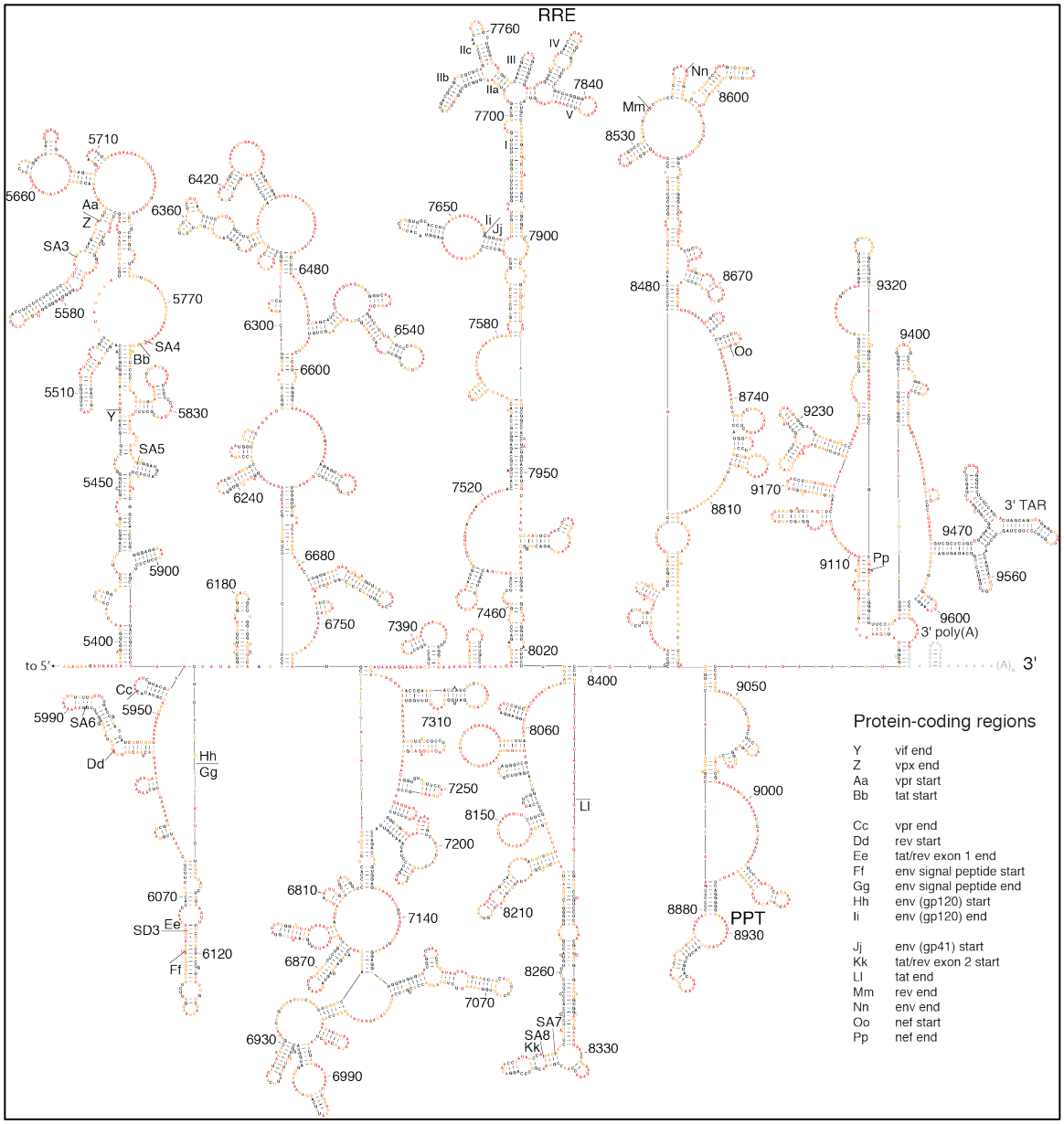


Figure 2.2: Secondary structure for the SIVmac239 genome with nucleotide identities shown. The structure is shown in two panels.


RRE
7760
IIc
IIb
III
IV
IIa
7840
7700
V
I
Nn
8600
Mm
8530
5710
5660
6420
Aa
Z
6360
7650
Ii
Jj
7900
SA3
6480
8480
8670
9320
5770
5580
6300
6540
7580
9400
SA4
Bb
6600
Oo
5510
8740
9230
Y
5830
SA5
7520
7950
5450
6240
9170
3' TAR
8810
9470
6680
9110
Pp
5900
9560
6180
7460
9600
6750
7390
3' poly(A)
5400
8020
3'
to 5'
8400
9050
Cc
5950
5990
SA6
7310
8060
Hh
Dd
Gg
7250
9000
8150
LI
7200
6070
6810
8210
8880
PPT
SD3
Ee
7140
8930
6120
Ff
6870
8260
7070
SA7
SA8
Kk
6930
8330
6990
Protein-coding regions
Y vif end
Z vpx end
Aa vpr start
Bb tat start
Cc vpr end
Dd rev start
Ee tat/rev exon 1 end
Ff env signal peptide start
Gg env signal peptide end
Hh env (gp120) start
Ii env (gp120) end
Jj env (gp41) start
Kk tat/rev exon 2 start
Ll tat end
Mm rev end
Nn env end
Oo nef start
Pp nef end

Region	Start	End	Nucleotides in Structure	GC	AU	GU
1*	1	488	1-539	99	60	16
2	720	804				
3#	2091	2193	2098-2187	11	14	4
4#	2434	2548	2462-2497	9	7	0
5#	2616	2871	2641-2895	32	33	10
6#	2969	3044	2996-3038	6	3	1
7#	3254	3361	3285-3323	5	5	0
8#	4679	4756	4679-4726	8	3	4
9#	4819	4914	4862-4907	5	6	1
10#	4926	5000	4976-4988; 5220-5232	7	4	2
11#	5409	5522	5392-5527	54	35	8
12#	5818	5987	5786-5946			
13#	7095	7236	7166-7223	6	8	2
14#	8200	8332	8233-8359	28	14	7
15	9433	9655				
G-P-P FS*	1793	1974	1793-1879	17	6	1
RRE*	7617	7853	7601-7909	47	25	13

Table 2.2: Start and end points corresponding to regions in the 75-nt moving window of median SIVmac239 SHAPE reactivities with values lower than 0.3 (from Figure 2b). Asterisks (*) indicate structures of known function and pound signs (#) indicate structures of unknown function used for GC versus AU/GU content and (G – A) analysis and comparison. For the GC versus AU/GU analysis, the valley at position 10, which corresponds to the 5' side of the longest continuous helix, was included along with the paired nucleotides on the 3' side. Nucleotides corresponding to valleys 11 and 12 were also taken together since they spanned both sides of the resultant stems. In all cases, we included 37 nts before and after each region to include all of the nucleotides that are included in the 75-nt window.

Highly structured regions in an RNA can be inferred in a model-free way by identifying regions with low overall median SHAPE reactivities. Many areas of the SIVmac239 RNA genome have low median SHAPE reactivity (defined as less than 0.3 on a scale from 0 to ~1.5) over a 75 nucleotide window, and these correspond to regions of structure with both known and unknown function (Figure 2.3). The lowest median SHAPE reactivity values occurred at the 5' and 3' ends of the genome. The highly structured 5' region extends until nucleotide 539 (Figures 2.1a and 2.3b, motif 1), and the structured 3' region begins at position 9462 at the start of the 3' TAR structure within the terminal repeat (R) regions (Figures 2.1b and 2.3b, motif 15). In addition, the Gag-Pro-Pol frameshift (G-P-P FS) element (Figures 2.1a and 2.3b; positions 1852-1879) and the RRE are highly structured (Figures 2.1b and 2.3b). By comparison, when we used RNA Decoder (a program that predicts evolutionarily conserved RNA secondary structure in the context of the protein-coding sequence of the RNA (137)) with an HIV-2/SIVsm sequence dataset to infer conserved regions of secondary structure, we found that the 5' and 3' UTRs and the RRE showed the strongest signal for conservation of structure (Figure 2.2). Using the RNA Decoder approach we conclude that major features of secondary structure are not conserved within the coding region at the level of the RRE. Other regions, however, have low median SHAPE reactivities, yet currently unknown RNA functions (Figures 2.1a, 2.1b, and 2.3b, motifs 2-14). In the following sections, we examine these structures and infer biological importance based on their conservation with HIV-1, first on a global scale then in detail.

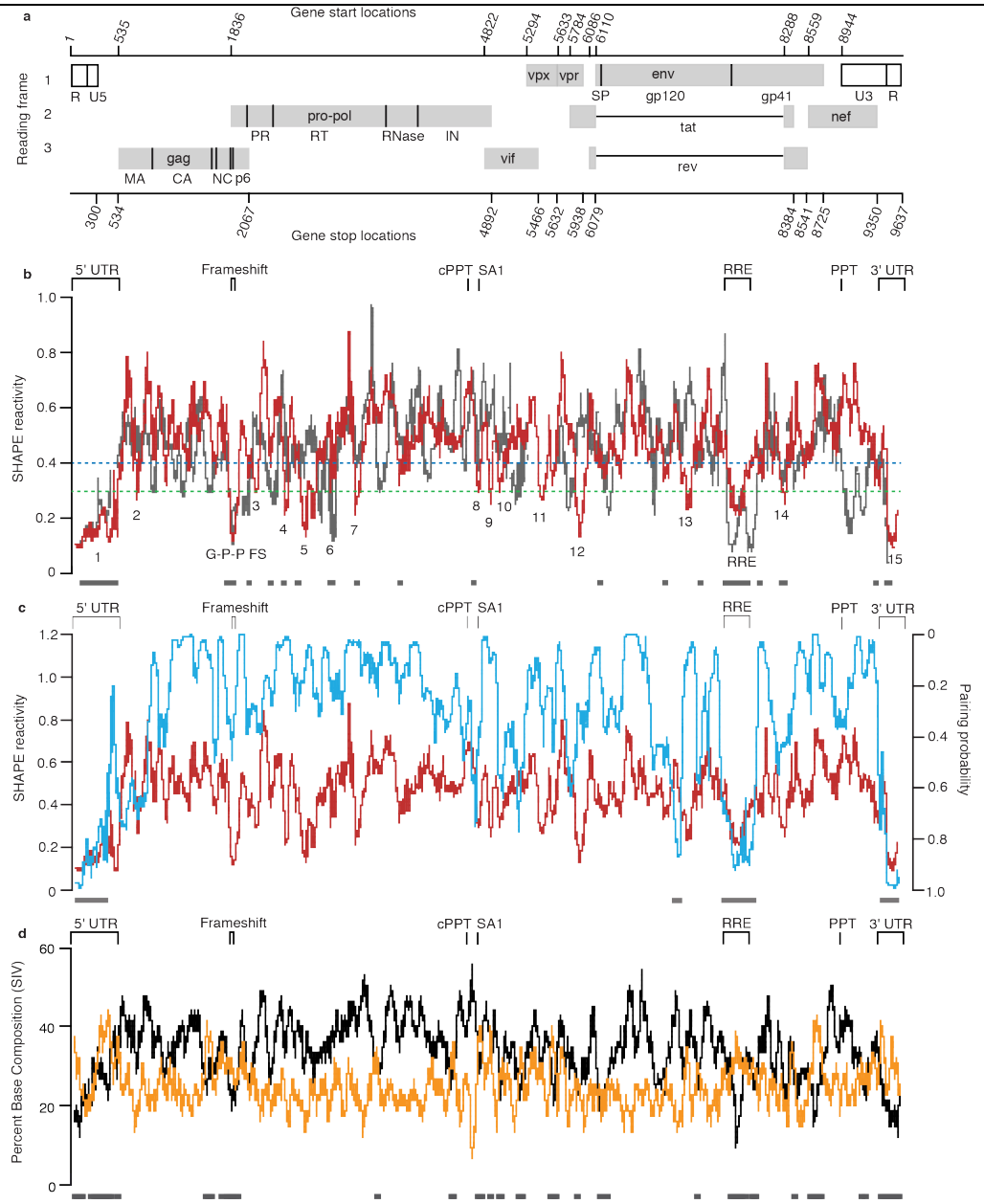


Figure 2.3: Genomic organization and SHAPE reactivity of SIVmac239 and comparison with HIV-1_{NL4-3}. (a) Organization of the SIVmac239 genome. Grey boxes indicate protein coding regions, dark lines indicate the boundaries of the mature viral proteins. (b) SIVmac239 median SHAPE reactivity values calculated over a 75 nucleotide sliding window (red). Green dashed line indicates SHAPE reactivity of 0.3. Regions with SHAPE reactivities below 0.3 are numbered (and listed explicity in Table 2.2). SHAPE reactivity values of HIV-1_{NL4-3} (gray) are shown as medians calculated over a 75 nucleotide sliding window and overlayed with those of SIVmac239. Viruses were codon-aligned based on the Los Alamos Database alignment (www.hiv.lanl.gov). Blue dashed line indicates SHAPE reactivity of 0.4, and the gray bars below indicate regions where median reactivity values for both viruses are below 0.4. (c) Comparison of SIVmac239 median SHAPE reactivity values (red) and median phylogenic pairing probability values (cyan). Pairing probabilities were calculated as described [3]. Genome regions where the median pairing probabilities are above 0.6 are indicated by gray. (d) Percent guanosine (gold) and adenosine (black) in the SIVmac239 genome over a 75 nucleotide sliding window. Gray bars below indicate regions where the percentage of guanosines is greater than the percentage of adenosines.

2. Overview of Base-pairing within the HIV-1_{NL4-3} and SIVmac239 RNA Genomes

SIVmac239 is the second full-length genomic primate lentivirus RNA evaluated by SHAPE-directed modeling; the first was that of HIV-1_{NL4-3} (172). Comparison of the structural models of these two distantly related retroviral RNA genomes should reveal conserved structural elements. For this analysis we have used updated folding parameters that result in modest changes in the previous HIV-1 model (see Methods). Visually the patterns of 1M7 reactivity in the 5' noncoding region, the frameshift site, and the RRE — all regions with well-established conserved functions — are similar for SIVmac239 and HIV-1 RNAs (Figure 2.3b). A bootstrapping analysis (see Methods) showed that the measured SHAPE profiles across both genomes were significantly more similar than expected by chance (10,000 trials, $p < 0.0001$). Thus, in a broad view, there appears to be a strong propensity to conserve the overall level of local RNA structure across the same regions of these two genomes.

The RNA folding algorithm employed for structure prediction included a pseudo-free energy change term to account for the SHAPE reactivity (see Methods). Newly optimized parameters for calculating the pseudo-free energy term were used to predict a revised secondary structure model for HIV-1_{NL4-3} based on the original reactivity data (172). We then compared the codon-aligned sequences of HIV-1_{NL4-3} and SIVmac239 for equivalency in terms of base-pairing. The two genomes share 50% identity at the nucleotide level; however, if these sequences were randomized, they would appear to have 24% identity, emphasizing the extent of divergence at 50% identity. We found that roughly half of the nucleotides predicted to be base-paired in the HIV-1_{NL4-3} sequence were also paired in the SIVmac239 sequence; conversely, only half of the nucleotides

predicted to be single-stranded in the HIV-1_{NL4-3} sequence were also single-stranded in the SIVmac239 sequence (Table 2.3). In spite of the limited conservation of paired bases, we did observe regions in similar locations within the genomes with low SHAPE reactivity (defined as median reactivity below 0.4 over a 75 nucleotide window). These areas (Figure 2.3b, grey dashes) largely fold into structures of unknown function. None of these structures have conserved base pairs, and in only a few examples are there even small hairpins that share pairing partners within 40 nucleotides in both alignments.

	Codon-Aligned Bases	Base-Paired	Base-Paired In Both HIV and SIV	Base-Paired With Same Partner	Base-Paired With Similar Partner (w/in 40 nts.)	Single-Stranded	Single-Stranded In Both HIV and SIV
SIV	8490	4970	2421	142	432	4671	2420
HIV		4500				4672	

Table 2.3: Comparison between SIVmac239 and HIV-1_{NL4-3} RNA genome secondary structure models. SHAPE-directed folding used ΔG_{SHAPE} parameters of m = 1.9 and b = -0.7.

As a more stringent definition of structural conservation, we tallied the number of base pairs in both genomes where both nucleotide positions of the base-pairing partners were maintained. Only 58 base pairs were fully conserved between the two genomes within the 8,738 nucleotides of the SIVmac239 coding region; an additional thirteen base pairs were conserved in the 5' UTRs. Overall, only 71 base pairs, 3% of the base pairs were precisely conserved in these two primate lentiviral genomes (Figure 2.1, blue bars in paired regions). Thus the regions that share a low SHAPE reactivity are areas of base pairing in both viruses, but the exact structures are not conserved between the two genomes.

3. Conserved Structure in the 5'-UTR

We aligned the 5'-UTRs of each virus both using the structures as predicted by SHAPE-directed modeling and by identifying the positions of the functionally conserved TAR region, 5' polyadenylation [poly(A)] signal, primer binding site (PBS), major splice donor (SD1) sequence, dimerization initiation sequence (DIS), and the Psi packaging element (Figure 2.4). Each of these functional elements folds into similar structures in both viral RNA genomes even though only thirteen of the ~180 predicted base pairs have the same two pairing partners in both 5' UTR regions (Figure 2.4, emphasized with heavy lines to indicate bonds). Additional conserved structures include a stem immediately 5' of the PBS, and the SL1 (DIS), SL2 (SD1), and SL3 helices. The stem containing the *gag* start site (Figure 2.4, MA start), initially identified by structure probing (36, 173), accounts for six of the identical pairing partners between HIV-1 and SIVmac239. This interaction has also recently been visualized by NMR analysis of the dimerized and

packaged form of HIV-1 RNA (100). There are structural differences in the TAR motif, which features three stem-loops in SIVmac239 but only a single stem in HIV-1. In addition, the stems in SIVmac239 5' UTR generally have more base pairs than the equivalent structures of HIV-1 (Figure 2.4).

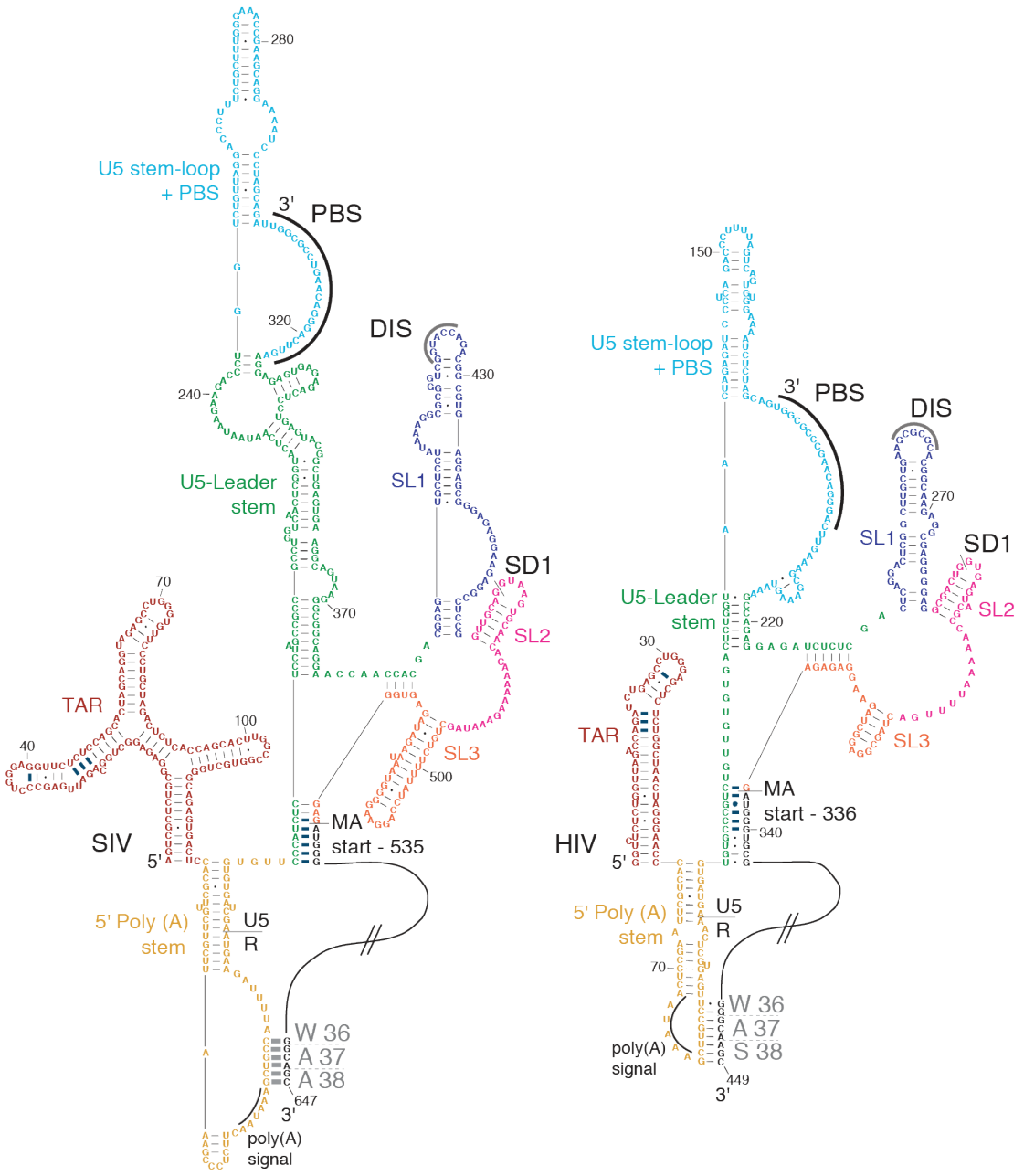


Figure 2.4: Structural similarity in the 5' regions of the SIVmac239 and HIV-1$_{NL4-3}$ genomes. Predicted structures of SIVmac239 (left) and HIV-1 (right) in the 5' region; distinctive structures are coded by color. Conserved base pairs are indicated by dark blue connecting lines. The primer binding site (PBS) and 5' poly(A) signal AAUAAA are emphasized with curved lines. The hatched lines indicate nucleotides within the MA coding domain that are not shown. The predicted pseudoknot in SIVmac239 is shown with thick gray lines; protein residues encoded by the pseudoknot region are shown.

Because of the sequence redundancies at each end of a retroviral genome, a poly(A) signal (AAUAAA) occurs at each end. The virus must prevent use of the signal at the 5' end to avoid producing truncated RNA transcripts. The 5' poly(A) signal lies in similar stem structures in both genomes, with unreactive loop nucleotides in the same area in SIVmac239 as in HIV-1. In the 5' poly(A) region of HIV-1, *in vitro* analysis suggests formation of a pseudoknot (134). The SIVmac239 SHAPE reactivity is low in the region in *gag* that corresponds, based on the codon alignment, to one of the pseudoknot stems in HIV-1. It is likely that a pseudoknot forms in this region of the SIVmac239 RNA as well (Figure 2.4). Conservation of the poly(A) stem-loop structure was previously noted for HIV-1, SIV, and HIV-2 sequences (134), whereas structural similarity in the MA region has only become apparent with SHAPE analysis (Figure 2.4). In sum, although only 13 base pairs are identical, the 5'-UTR structure is highly conserved between the SIVmac239 and HIV-1 RNA genomes both at the level of overall structure (Figure 2.3b) and in terms of the local architecture of multiple functional elements (Figure 2.4).

4. Conserved Structure in the Gag-Pro-Pol Frameshift Region

The primate lentiviruses generate more Gag protein than the Gag-Pro-Pol product via a minus-one frameshifting process. Frameshifting occurs at a "slippery sequence", a poly(U) region, and is potentiated by a downstream structure that stalls the ribosome (88, 136). The RNA structure in the region of the frameshift site is similar in HIV-1_{NL4-3} and SIVmac239. In both cases, the poly(U) slippery sequence is part of a stem, although the poly(U) region is not paired in the SIVmac239 structure. In addition, there is a second

stem just downstream of the U stretch; however, the frameshift stem is further from the poly(U) sequence in SIVmac239 than it is in HIV-1_{NL4-3} (Figure 2.5). Despite the fact that this region carries out a conserved and essential function in retrovirus replication, the organization of these stems is different in the two genomes and they have no shared base pairs (Figure 2.5).

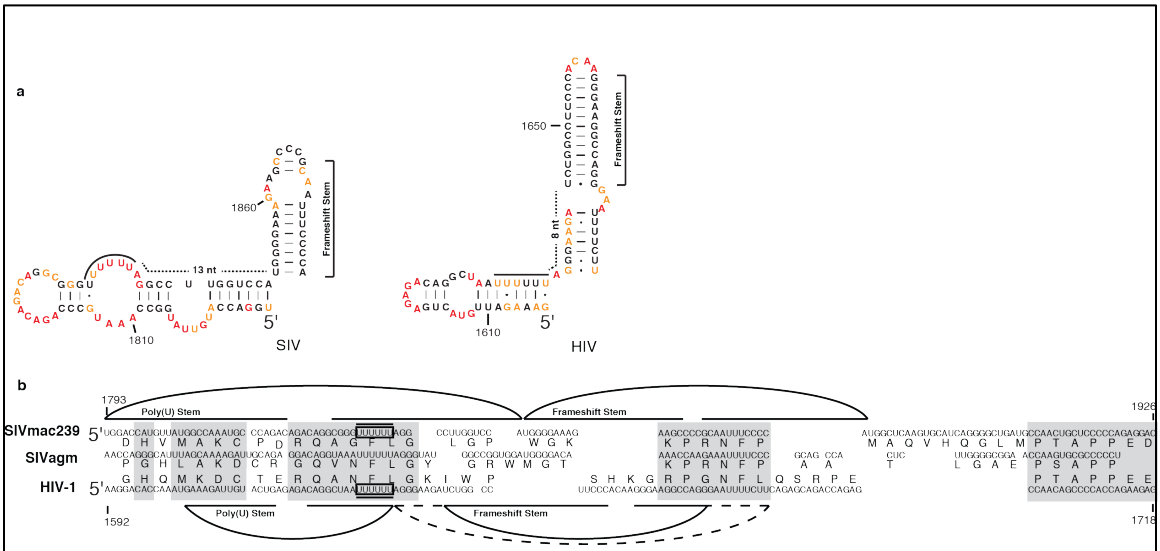


Figure 2.5: Codon alignment and predicted pairing partners in the Gag-Pro-Pol frameshift region of SIVmac239 and HIV-1. (a) RNA structures at the Gag-Pro-Pol frameshift stem for SIVmac239 (left) and HIV-1 (right). The main stem is emphasized by brackets and the poly(U) slippery sequence is emphasized by a curved line. The dotted line indicates nucleotides between the poly(U) sequence and the beginning of the frameshift stem. (b) The sequences of HIV-1, SIVagm, and SIVmac239 were aligned horizontally. The poly(U) slippery sequence is boxed. Curved lines represent base pairs between the nucleotides within HIV-1 and SIVmac239; the curved lines connecting nucleotides downstream of the poly(U) sequence correspond to the frameshift stems, the dotted line represents the extended frameshift stem in HIV-1. Gray boxes indicate regions of strong alignment, and spaces in the sequence indicate regions of poor alignment. The structure of the HIV-1 hairpin is modified from the previous model(172) based on the updated folding parameters as described in the Methods.

We attempted to define the pathway through which these structures evolved by examining the sequences surrounding the poly(U) slippery sequence. To facilitate this analysis, we included a sequence from SIVagm (GenBank accession M30931), which is distantly related to both SIVmac239 and HIV-1_{NL4-3}. All three sequences aligned well at the protein and nucleotide levels upstream and through the poly(U) slippery sequence (Figure 2.5, gray boxes). However, this alignment is lost three nucleotides 3' of the poly(U) sequence. Sequence was again aligned at the conserved PTAPP motif in Gag (Figure 2.5, 3'-most gray box). One possible explanation for the abrupt loss of sequence alignment in this region is that the frameshift hairpin itself in mutagenic, consistent with the idea that structure in the RNA would induce pausing which enhances recombination and mutation during viral DNA synthesis (54, 92). These hairpins are predicted to be among the most stable in each genome. A structure stable enough to stall the ribosome is likely also to induce pausing of reverse transcriptase during DNA synthesis, increasing the possibility of recombination in the region of the frameshift hairpin. Thus, we hypothesize that the rapid evolution of structure in this region is due to the mutagenic effect of the structure itself.

5. Conserved Structure in the Rev Response Element (RRE)

The RRE includes binding sites that mediate oligomerization of the Rev protein; oligomerized Rev mediates export of unspliced and singly-spliced viral RNA from the nucleus (110). The sequence is conserved in this region of many primate lentivirus genomes (94). The predicted RRE structure (Figure 2.1) consists of a long, irregular stem I helix terminated by a set of small hairpins including the IIb stem, previously described

as the primary Rev binding site (30, 83), and the auxiliary hairpins (stems III, IV, and V) that facilitate multimerization of Rev (39). Twenty-nine of the 71 base pairs conserved between HIV-1_{NL4-3} and SIVmac239 are in the small terminal hairpins in the RRE (Figure 2.6, blue bars); these nucleotides are 78% conserved at the sequence level. By contrast, the long stem is mostly devoid of conserved pairing partners, and there is a shift by one nucleotide in the codon-aligned pairing. When we used the codon alignment to force superposition of the SIVmac239 sequence onto the RRE structure of HIV-1_{NL4-3} (contrary to the SHAPE-directed structure), there was a large reduction (30%) in the number of base pairs formed in stem I (Figure 2.6). Thus, conservation of specific base pairs is limited to four of the small hairpins that serve as protein interaction regions, whereas the long stem has pairs that are shifted by one nucleotide (Figure 2.6). We infer that neither the sequence nor the exact base-pairing partners of the long stem I are critical, although its presence and long length are conserved.

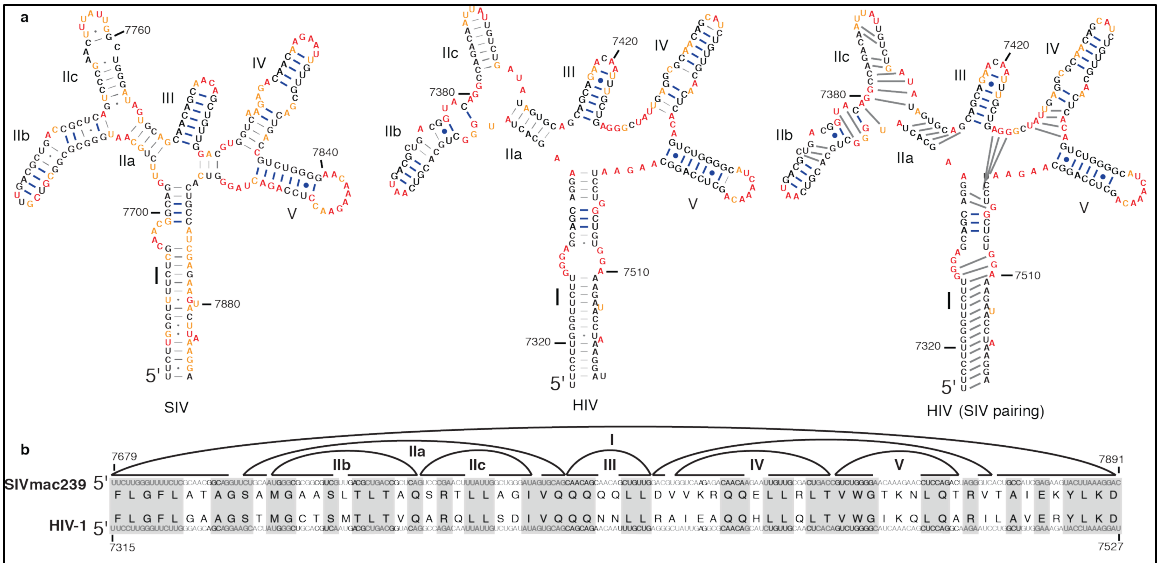


Figure 2.6: Codon alignment and predicted pairing partners in the RRE region of SIVmac239 and HIV-1. (a) Predicted structures within the RRE are shown for SIVmac239 (left) and HIV-1 (middle). The HIV-1 structure with forced SIVmac239 pairing is also shown (right). Blue lines indicate base pairs that are exactly conserved between the two viruses. (b) The sequences of SIVmac239 (top) and HIV-1_{NL4-3} (bottom) aligned horizontally. Curved lines indicate base-pairing partners. Gray boxes indicate regions of amino acid alignment. Roman numerals indicate helices discussed in the text.

6. Conserved Structure at the First Splice Acceptor Site SA1

Retroviruses use diverse splice donor (SD) and splice acceptor (SA) sites to generate a multitude of spliced mRNAs which direct synthesis of the small regulatory proteins and Env while retaining some unspliced RNA for both translation of Gag and Gag-Pro-Pol and for packaging of the full-length genome into new virions (142). Splicing to generate these mRNAs is highly regulated. This regulation takes place at both the sequence level and at the RNA structure level. Five of the base pairs that are precisely conserved between HIV-1_{NL4-3} and SIVmac239 are in the stem of the hairpin structure that contains the first splice acceptor site (SA1) (Figure 2.7), which is used to generate the transcript that codes for the viral protein Vif (165). Most of the other splice acceptor regions (SA2-SA8) downstream of SA1 are part of short hairpins as well, with the exception of SA4 (Figure 2.1). Each of these short hairpins has low median SHAPE reactivity (most are below the overall median of 0.46); however, only the hairpin at the SA1 site is exactly conserved between HIV-1 and SIVmac239 with several of the same pairing partners. The viral splice acceptors have weak splicing sequences to allow balanced usage of each with the major splice donor SD1 (129). The relative strengths of HIV-1 splice acceptor sequences have been previously analyzed, and splicing is most efficient to SA1 (76). We propose that the conserved stem-loop structure at SA1 down-regulates splicing to this site to ensure sufficient use of the other downstream splice acceptor sites.

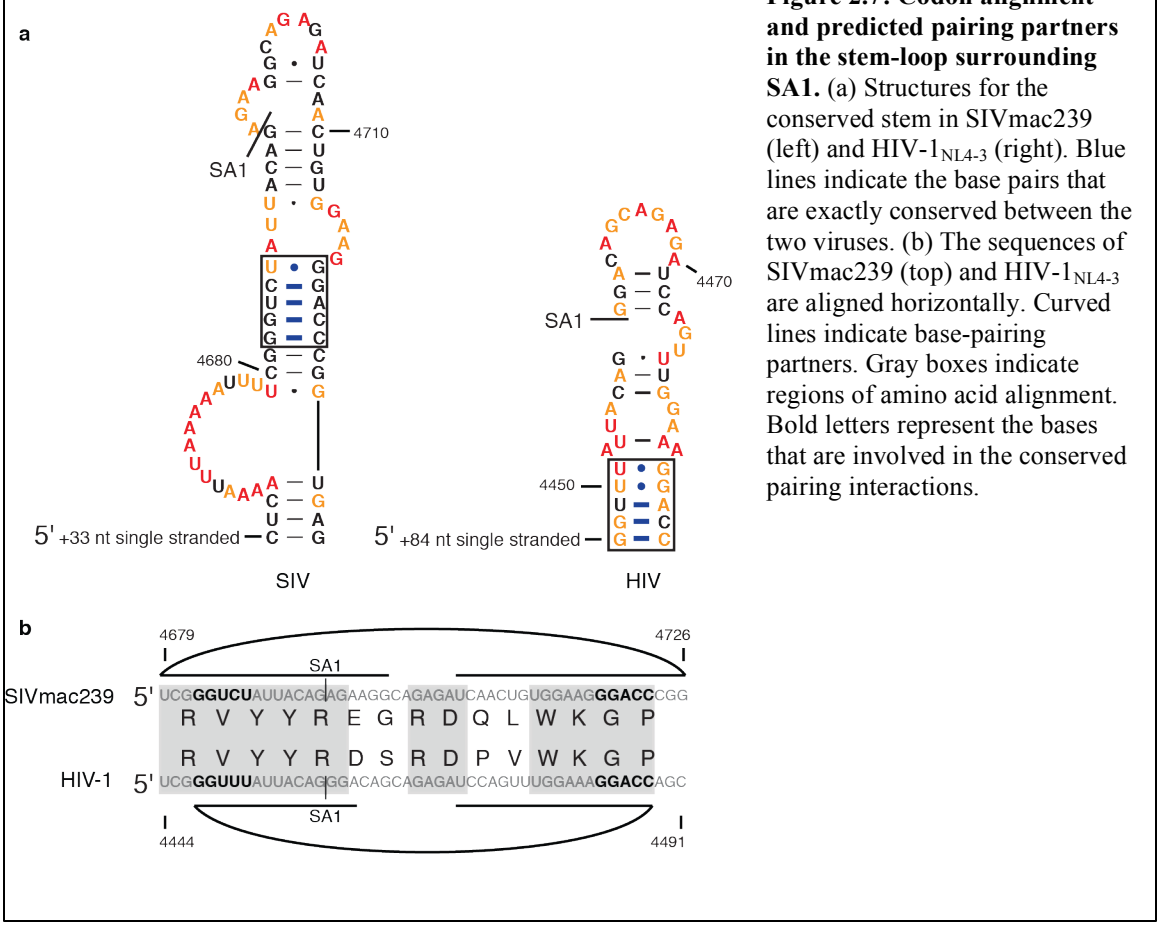


Figure 2.7: Codon alignment and predicted pairing partners in the stem-loop surrounding SA1. (a) Structures for the conserved stem in SIVmac239 (left) and HIV-1_{NL4-3} (right). Blue lines indicate the base pairs that are exactly conserved between the two viruses. (b) The sequences of SIVmac239 (top) and HIV-1_{NL4-3} are aligned horizontally. Curved lines indicate base-pairing partners. Gray boxes indicate regions of amino acid alignment. Bold letters represent the bases that are involved in the conserved pairing interactions.

7. Conserved Structure in the cPPT and PPT

Retroviruses prime plus-strand DNA synthesis from polypurine tracts (PPT) that are derived from viral RNA during minus-strand DNA synthesis (161). These primers are resistant to degradation by RNase H, and their specificity as second-strand primers is enhanced by the viral NC protein (141). Primate lentiviruses prime from two regions, one near the center of the genome (cPPT) and one just upstream of the U3 sequence near the 3' end of the genome (PPT) (23). We observed a common structural motif in these polypurine tracts in SIVmac239 and in HIV-1. cPPT and PPT motifs in both viruses contain a 5' A-rich single-stranded region followed by a 3' G-rich base-paired region (Figure 2.8a) that have strikingly similar patterns of SHAPE reactivity (Figure 2.8b). Since the PPT and cPPT function as second-strand primers while hybridized with the first/minus strand of viral DNA, it is unlikely that RNA secondary structure *per se* is relevant to the function of these sequences as plus strand primers. These patterns drew our attention to the possibility that guanosine and adenosine play very different roles in defining secondary structure and that these roles might be reflected in the structures of the PPTs in genomic RNA but as a byproduct of their high G content in the context of a purine-rich run. This caused us to consider the role of base composition in defining secondary structure more broadly across the genome.

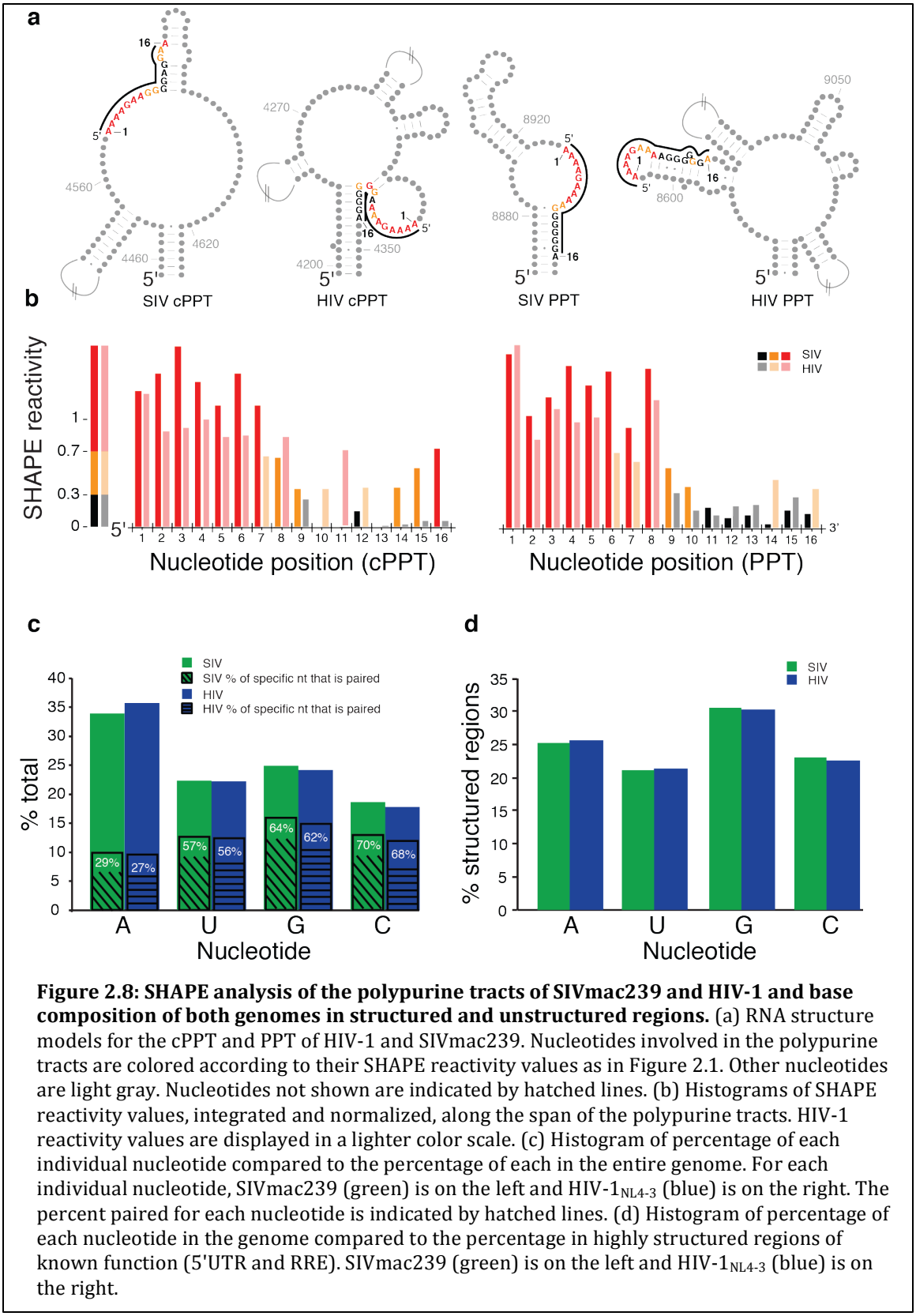


Figure 2.8: SHAPE analysis of the polypurine tracts of SIVmac239 and HIV-1 and base composition of both genomes in structured and unstructured regions. (a) RNA structure models for the cPPT and PPT of HIV-1 and SIVmac239. Nucleotides involved in the polypurine tracts are colored according to their SHAPE reactivity values as in Figure 2.1. Other nucleotides are light gray. Nucleotides not shown are indicated by hatched lines. (b) Histograms of SHAPE reactivity values, integrated and normalized, along the span of the polypurine tracts. HIV-1 reactivity values are displayed in a lighter color scale. (c) Histogram of percentage of each individual nucleotide compared to the percentage of each in the entire genome. For each individual nucleotide, SIVmac239 (green) is on the left and HIV-1_{NL4-3} (blue) is on the right. The percent paired for each nucleotide is indicated by hatched lines. (d) Histogram of percentage of each nucleotide in the genome compared to the percentage in highly structured regions of known function (5'UTR and RRE). SIVmac239 (green) is on the left and HIV-1_{NL4-3} (blue) is on the right.

8. The Role of Base Composition in Defining Structure

The base compositions of both the HIV-1 and SIVmac239 RNA genomes are dominated by adenosine (34%); the percentage of cytidine is low, around 17%, and the percentages of guanosine and uridine are each about 25% (Figure 2.8c). Regions with large numbers of base pairs must have approximately the same number of pyrimidines and purines. In structured regions with known function, including the 5' UTR and RRE in both HIV-1 and SIVmac239, the average base composition is 25% A, 29% G, 22% C, and 22% U (Figure 2.8d). The higher percentage of guanosines compared to adenosines in these base-paired regions has the effect of concentrating more adenosines in unpaired regions, where adenosines represent fully half of the nucleotides. Only about 30% of the adenosines are base-paired, whereas approximately 60-70% of guanosines, cytidines, and uridines are base-paired in the two lentivirus models (Figure 2.8c). This trend toward favoring unpaired adenosines but paired guanosines, cytidines, and uridines is also observed in other highly structured regions of RNAs, including bacterial ribosomal RNAs (60). However, given the A-rich primate lentiviral genome the overall effect is to create A-rich single-stranded regions.

We considered the possibility that the greater stability of the G-C base pair relative to A-U or G-U base pairs might define structures that are conserved between the two genomes. We compared the base compositions of structures within SIVmac239 with known function (5' UTR, RRE, and the frameshift stem) to regions of extensive structure but unknown function. The conserved structures with known function have a higher average guanosine content and a significantly higher ($p = 0.04$) percentage of G-C pairs (57.9%) than structured regions of the genome that are not conserved (49.5%) (Table

2.2). In support of this idea, we noted that the SHAPE reactivity was higher for both adenosine and guanosine residues in regions of non-conserved structures (data not shown), which suggests that the adenosine-rich structures may allow for unfolding and refolding to occur more readily or may allow for the presence of multiple conformations compared to guanosine-rich structures in functional regions. We also hypothesize that selection to maintain functional secondary structures in primate lentiviruses such as HIV-1_{NL4-3} and SIVmac239 has resulted in regions defined by clusters of guanosines within an otherwise adenosine-rich genome. In Figure 2.3d we show an analysis of A versus G content across the SIVmac239 genome using a sliding window of 75 nucleotides. For several of the structures with known function there is a reciprocal relationship with an increase in G content and a decrease in A content, consistent with the idea that there is selective pressure to maintain islands of high G content to anchor structure.

C. Summary

Ultimately, HIV-1 and SIVsm/HIV-2 genomic RNAs accomplish many of the same functions in the context of viral replication. There is abundant evidence that RNA structure is either critical or directly modulates functions including viral DNA synthesis, RNA splicing, genome packaging, and mediating interactions with both viral and cellular proteins. One paradigm for assessment of conserved function is that of the ribosomal RNAs where strong base-pairing patterns are highly conserved despite large sequence variations over the course of evolution. We sought to identify functionally important RNA secondary structures in the primate lentivirus genome by comparison of SHAPE-directed nucleotide-resolution structure probing information and by developing structural

models of representative HIV-1 and SIVsm genomes. We developed a secondary structure model for SIVmac239 and compared it to a modestly revised structural model for the HIV-1_{NL4-3} genome (172). These genomes share about 50% sequence identity, and, although a similar fraction of each genome is base-paired (60%), only about 3% of predicted base pairs were with identical partners within the coding region of the genome. Almost one-half of these identical pairs were clustered in the Rev binding domain of the RRE. Thus, there has been massive reorganization of the patterns of RNA secondary structure between these two genomes.

Even within regions of highly conserved function, there were large differences in the sequence and pairing partners. Dramatic differences in the structure of TAR, longer stems in the 5' UTR of SIVmac239 relative to that of HIV-1, different pairing partners and poor sequence alignment in the Gag-Pro-Pol frameshift stem, and a one-base shift in the alignment in the RRE stem all point to remodeling of these domains. It has been shown that elements of secondary structure can promote recombination during retroviral DNA synthesis (54, 92). Certain stable structural elements in fact appear to be mutagenic. In regions like the Gag-Pro-Pol frameshift stem, selective pressure does not maintain a particular set of base pairs, but rather ensures that a sufficiently stable structure exists. In contrast, regions involved in RNA-protein interactions, such as the Rev oligomerization domain, displayed a significantly higher level of conservation than other regions of the genome, indicating that selective pressure maintains a particular structure for interaction with protein.

In the regions of conserved secondary structure, we observed significantly higher guanosine content compared to the overall base composition of the genome. Higher

levels of guanosine content may function to stabilize functionally critical structures. The lentivirus genomes are adenosine-rich, and the resulting less stable secondary structures are likely to exist in one or more alternative states, even if they are drawn as a single representative structure in our models. We propose that scanning for guanosine-rich regions in these and other adenosine-rich viral genomes may help identify important structural domains. One source of the selective pressure to maintain an adenosine-rich genome is the action of APOBEC3-G and -F, enzymes that deaminate cytidines on the DNA minus strand during viral DNA synthesis giving rise to G-to-A transitions on the plus strand (14). Although these lentiviral genomes are adenosine-rich, they are not guanosine-poor (approximately 25% G content) but rather cytidine-poor (at 17% C content). Thus mechanisms must be in place to retain guanosines in these regions of functional RNA secondary structure.

Our analysis shows that within the short evolutionary distance between the HIV-2/SIVsm lineage and HIV-1, the primary features of secondary structures at the individual base pair level are at the ends of the RNA and in the RRE. This observation is consistent with the interpretation that, for most of the lentivirus genome, there is little selective pressure to maintain specific pairing interactions. This is in contrast to the evolutionary pressure on ribosomal RNA. The sequences of 16S rRNA are less than 50% conserved when eubacterial, archaebacterial, and eukaryotic RNAs are compared, but the structure has been maintained through evolution by mutations that compensate for changes in the sequence directly affecting the base pairing (reviewed in (128)). In strong contrast, the lentiviral RRE structure, particularly in Stem I, did not evolve through base changes that maintained pairing. We previously examined a model where higher rate of

transition versus transversion mutations exist in paired regions of many RNA structures (presumably to maintain pairing partners), but found that the HIV-1 RRE was the exception as its mutation pattern did not fit this model (86). The relatively low conservation of base pairs between HIV-1_{NL4-3} and SIVmac239 is consistent with this observation since very few pairing partners are maintained.

We propose that the lentiviral genomic structure is evolving in the context of two significant mutagens. APOBEC3-G and -F indirectly mutate guanosines to adenosines, which weakens stability of structural motifs. The structural motifs themselves are mutagenic during DNA synthesis. The effect of these mutagens is filtered by the selective pressure to maintain useful structural motifs. The majority of the genome, in contrast, is depleted of both guanosines and strong secondary structure, and thus has evolved to be less susceptible to these mutagens.

C. Materials and Methods

1. Virus production

An infectious clone of SIVmac239 (GenBank accession M33262) was a gift from Ronald C. Desrosiers (New England Regional Primate Center, Harvard Medical School)(81). SIVmac239 was used to infect SupT1 CCR5 CL.30 cells (a gift from J. Hoxie, University of Pennsylvania); these cells are a non-Hodgkin's T cell lymphoma cell line (a modified version of the SupT1 cell line) (117). The virus produced was purified as described (25).

2. Genomic RNA

Viral genomic RNA was extracted from purified SIVmac239 viral particles as described (172) in a manner that avoided denaturation of RNA secondary or tertiary structure. Viral RNA was extracted using phenol/chloroform after lysis and treatment with Proteinase K. No heating steps, chelating agents, or chemical denaturants were used in the purification. The final RNA product was precipitated in 70% (v/v) ethanol with 300 mM NaCl and stored at -80 °C until use.

3. SHAPE analysis of RNA

RNA was treated as described (172). Briefly, the precipitated RNA was collected by centrifugation and the ethanol removed. Each pellet, containing 62 pmol of SIVmac239 genomic RNA, was individually resuspended in 620 µl of 50 mM HEPES (pH 8.0), 200 mM potassium acetate (pH 8.0), 3 mM $MgCl_2$ and incubated at 22 °C for 10 min then at 37 °C for 15 min. Aliquots of 32 µl of 45 mM 1M7 in dimethyl sulfoxide (DMSO) (122) or DMSO alone were warmed at 37 °C for 30 sec, then 288 µl of the RNA solution was added to each and incubated at 37 °C for 5 min. RNA was recovered by adding 32 µl of 50 mM EDTA (pH 8.0) and precipitation with ethanol.

4. Primers

Each primer contained a 5' six carbon linker terminated with an amino group (IDT); a total of 38 primers were used (Table 2.1). The primers were tethered to 5-FAM or 6-JOE fluorophores (AnaSpec) using N-hydroxysuccinimide chemistry. Purified primers were

spectrophotometrically determined to have at least 82% labeling efficiency, with most labeled to greater than 95%, as determined by the [dye]/[DNA] ratio.

5. Primer extension

Both the (+) and (–) 1M7 reagent reactions were subjected to reverse transcription with FAM-labeled primers using SuperScript III Reverse Transcriptase (Invitrogen). A sequencing length ladder was generated using the JOE-labeled primers and termination with a dideoxynucleotide. After cDNA synthesis, the reverse transcription reaction products were combined with their corresponding JOE-labeled sequencing reactions, the latter performed using plasmids containing SIVmac239 sequences, p239SpSp5', and p239SpE3' (obtained through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH from Dr. Ronald Desrosiers (81)). Primer extension products were resolved by length using an Applied Biosystems AB3130 capillary electrophoresis instrument.

6. Data processing

ShapeFinder software (http://bioinfo.unc.edu) was used to convert the raw capillary electrophoresis electropherograms of fluorescence intensity to normalized SHAPE reactivities (40, 164, 173). Data were processed as described (172).

7. SHAPE-directed RNA structure modeling

Inclusion of SHAPE information provides an experimental adjustment to the well-established nearest neighbor model for RNA folding (114). For secondary structure

prediction, SHAPE data are incorporated as a pseudo-free energy change term, ΔG_{SHAPE}, implemented in *RNAstructure* (40):

$$\Delta G_{SHAPE} = m\ ln[\text{SHAPE} + 1] + b \quad (1)$$

The slope, m, corresponds to a penalty for base pairing that increases with the experimental SHAPE reactivity, and the intercept, b, reflects a favorable pseudo-free energy change term for base pairing at nucleotides with low SHAPE reactivities. These two parameters must be determined empirically. When Watts et al. analyzed the HIV-1_{NL4-3} genome, m = 3.0 and b = -0.6 were the optimal parameters (172) and, in general, these parameters still perform well. The current, recently updated, parameters are m = 1.9 and b = -0.7 give the highest sensitivities in a bootstrapping statistical analysis of multiple RNAs (C.E. Hajdin, personal communication). Changing the slope and intercept parameters from m = 3.0, b = -0.6 to m = 1.9, b = -0.7 results in a reduction of 34% of the specific base pairs that were predicted in the HIV-1_{NL4-3} genome, most of which are in weakly structured regions.

8. RNA secondary structure model

The SIVmac239 sequence (9646 nucleotides) with the addition of a poly(A) tail consisting of 10 adenosines was folded using the *RNAstructure* algorithm (114, 146). SHAPE reactivities were incorporated into the thermodynamic folding algorithm to constrain secondary structure. Due to computational restrictions in the folding algorithm (caused by the length of the genomes), folding was accomplished in large overlapping pieces consisting of at least two-thirds of the entire genome. The structure of the whole genome was generated by combining the separately folded pieces at identical structures.

Multiple analyses with varying lengths gave consistent structures. Although we are unable to identify pseudoknots *de novo* with the current algorithm, the model includes the pseudoknot at the 5' poly(A) stem that we predict based on low SHAPE reactivity of nucleotides in loop regions and by sequence alignment with HIV-1_{NL4-3}. Recently updated folding parameters m = 1.9 and b = -0.7 (C.E. Hajdin, personal communication) were used to generate a new version of the HIV-1_{NL4-3} RNA structure, which was used for these analyses.

9. Sequence alignment

HIV-1_{NL4-3} and SIVmac239 sequences were aligned at the codon level using the Los Alamos lentivirus compendium (www.hiv.lanl.gov). Protein start and end positions, known RNA structures, and known protein functional regions were taken into consideration as well as conserved amino acids. Deletions and insertions were incorporated into the sequence alignment where appropriate.

10. Statistical analyses

A Matlab 7.8 (R2009a) script was used to compare the actual average absolute difference in SHAPE reactivity value across all aligned HIV-1_{NL4-3} and SIVmac239 genome positions. We randomized the position assignments for reactivity values in the SIVmac239 genome to make a distribution of average absolute differences, then repeated this 100,000 times to generate a random distribution curve and plotted the observed average on this curve. We employed a two-tailed Fishers exact test to compare the GC

content to AU and GU content of structures with known and unknown function in SIVmac239.

11. RNA structure display

The secondary structures of the RNA models were organized using xrna (http://rna.ucsc.edu/rnacenter/xrna).

12. Grammar predictions of structure

Structure predictions using RNA-Decoder (137) were performed as previously described (172) with the following modifications. The input alignment was a reduced version of the HIV-2 web alignment available from the Los Alamos lentivirus compendium (www.hiv.lanl.gov). Codon positions in overlapping regions were designated according to the reading frame of the first member of the following pairs: gag-pro, pol-vif, vif-vpx, vpr-tat1, tat1-rev1, env-tat2, env-rev2, env-nef. The alignment was scanned using separate phylogenetic trees for the upstream and downstream sections, which were generated by Tree-Puzzle[6] using the GTR+γ(4) model, 10,000 puzzling steps, "accurate" parameter estimation, and other default settings. The tree for the first half of the genome was built on the third codon positions of the gag, pro, pol, and vif genes and the 5' non-coding region, and the downstream tree was inferred from the third positions of the vpx, vpr, tat1, env, and nef genes and the 3' non-coding region. Trees are available from the authors on request.

CHAPTER 3

THE EFFECT OF RNA SECONDARY STRUCTURE ON SPLICING REGULATION AT THE 3' SPLICE SITE SA1 AND ANALYSIS OF REGULATORY STRUCTURES IN HIV-1 *IN VITRO* mRNA TRANSCRIPTS[2]

A. Introduction

Lentiviruses preserve their complex transcriptome while still maintaining the size of their relatively small RNA genome by utilizing the host splicing machinery. The human immunodeficiency virus type 1 (HIV-1) RNA primary transcript contains four splice donor sites (5'ss, SD1-4) and eight splice acceptor sites (3'ss, SA1-7), which are combined to produce more than 40 mRNA species (142). Each transcript exits the nucleus in one of three states: unspliced, singly spliced, or multiply spliced. The fate of unspliced RNA transcribed from the integrated proviral DNA is either to dimerize and be packaged with the budding virus (73, 91, 121, 133, 140) or to be used as a template for Gag and Gag-Pro-Pol translation. Splicing of this viral RNA, however, results in an elaborate pattern of transcribed messages that allows for increased diversity of viral proteins (142). RNAs that are singly spliced yield the longer 4kb class of transcripts representing the mRNAs for the Vif, Vpr, Vpu, and Env proteins. The shorter 1.8 kb class of multiply spliced transcripts represent the mRNAs for the Tat, Rev, and Nef proteins (142).

---

[2] Megan Wise contributed to the production of plasmids for use as templates for *in vitro* transcription.

Viral RNA splicing is controlled to ensure that all needed RNA variants are generated, including a fraction that remains completely unspliced. For RNAs that are spliced, the major splice donor (SD1) 5'ss is used in all splicing events, and multiple splice acceptors are available. The *tat* transcript is formed when 3'ss SA3 is used. The *rev* transcript can result if one of three 3'ss (SA4a, SA4b, or SA4c) is used. When 3'ss SA5 is used, the *nef* transcript can result. Fully-spliced messages subsequently excise the intron between the 5'ss SD4 and the 3'ss SA7, yielding the 1.8 kb class. If 3'ss SA5 is used with no other downstream splicing, the result is the *vpu*/*env* transcript. The other two transcripts without downstream splicing are 3'ss SA1 and SA2 which are the *vif* and *vpr* messages, respectively (142). If any of the splice acceptors were too strong, they would be preferentially utilized. Instead, the HIV-1 genome contains a series of weak splice donors and acceptors coupled with weak branch-point sequences throughout the RNA (4, 46, 129, 153, 157). In this way, the usage of each site is controlled to produce the proper amount of all the necessary transcripts.

Along with weak sequences, the splicing machinery is further controlled by multiple *cis* regulatory elements within the sequence and structure of the RNA (Figure 3.1) (Reviewed in (160)). RNA structure has been implicated in regulation of splicing at the 5'ss SD1 whose stem loop must be kept at a certain stability for proper usage of the major donor site (2). Splicing of the *vpr* transcript occurring at 3'ss SA2 is regulated by an exonic splicing silencer (ESSV) downstream of SA2, which folds into a hairpin and binds heterogeneous nuclear ribonucleoprotein A1 (hnRNP A1) in the loop region (103, 149). The 3'ss SA3 which is used for splicing of the *tat* message is regulated in multiple ways by a long asymmetrical stem-loop including SLS3 that contains an exonic splicing

enhancer sequence (ESE2) which binds elements including SR protein SC35 to increase splicing at 3'ss SA3 (61). This splicing enhancement is counteracted by both an exonic splicing silencer (ESS2p) in a short stem-loop SLS2 near SA3 that controls the usage of the splice acceptor by binding hnRNP H, as well as the sequence ESS2 on SLS3 which acts by binding hnRNP A/E (3, 70, 71, 176). Regulation at SA4c/a/b and SA5 along with the 5'ss SD4 is accomplished by a guanosine-adenosine-rich ESE called GAR that binds SR proteins to upregulate splicing events in this region (20, 75). The structure around the GAR sequence has not yet been described as playing a regulatory role. Usage of the 3'ss SA7, which is necessary for removal of the intron for the *tat*/*rev*/*nef* mRNAs, is tightly regulated by RNA structures SLS1 including the ISS element and a SLS3 including the bipartite ESS3 element which cooperatively bind hnRNP A/B proteins to yield unspliced and incompletely spliced transcripts for *vif, vpr,* and *vpu*/*env* mRNA, while another stem loop (also termed SLS2) contains the Janus element ESE3 which is the initiation site for hnRNP A1 binding, yet also is able to bind SR proteins SF2/ASF, effectively both enhancing and silencing splicing at SA7 (37, 111). RNA structure has been implicated in either hiding or exposing various regulatory elements to enhance or seclude protein-binding efficiency to these sequences (reviewed in (169)). In this way, RNA structure plays a substantial role in splicing regulation throughout the virus.

The lentiviral Vif protein is important in the downregulation of APOBEC3-G and –F cellular restriction factors (78, 112, 151, 175). The transcript for Vif is a singly spliced product that results when the major 5'ss SD1 is joined to the 3'ss SA1, but none of the other possible downstream splice donors are used (SD2, SD3, or SD4) (142). Splicing control that reduces the production of the *vif* mRNA has been described wherein a

suboptimal donor sequence and a GGGG silencing site at the downstream 5' ss SD2 along with upstream sequences ESEVif, ESEM1, and ESEM2 regulate transcription of the full *vif* message (48, 104, 107). The sequence at SA1, including these downstream elements, has been characterized as one of the strongest acceptor sequences, (76, 107). None of these studies, however, have investigated the structure at SA1 or its effect on splicing regulation. The full-length genomic HIV-1 RNA has recently been modeled using the selective 2'hydroxyl acylation analyzed by primer extension (SHAPE) RNA structural probing method (40, 45, 56, 72, 173), identifying a hairpin structure around the 3'ss SA1 (172). The base pairs that interacted to form the stem-loop structure at the SA1 site were 5 of only 71 conserved pairs in a SHAPE analysis of simian immunodeficiency virus (SIVmac239) RNA genomic structure. The conservation of structure at this site in an otherwise rapidly evolving genome suggests a function for this RNA stem-loop structure, perhaps in splicing regulation (Pollom *et al*., submitted).

The availability of two full-length lentiviral genomic RNA structures makes identification of conserved regions possible. We tested the function of the evolutionarily conserved stem-loop at the 3'ss SA1 (which we term SLSA1) to determine its role in viral replication and splicing regulation and demonstrate that mutations to SLSA1 affect the fitness and splicing profile of the virus. In an effort to understand how the removal of various introns in the production HIV-1 mRNAs affects their structure, we used SHAPE to determine the RNA secondary structures of the most abundant spliced RNA variants. These structures have previously only been analyzed in small segments of RNA or by computational methods. Here, we identify the structural context of the splicing regulatory elements that remain present within the various forms of the HIV-1 mRNA transcripts.

Figure 3.1: HIV-1 splicing regulatory sequences. Sequences that have been indicated in regulation of splicing the HIV-1 pre-mRNA are shown as well as sites of splice acceptors (SA1-7) and donors (SD1-4). Arrows indicate positive regulation (enhancement) toward the directed splice site and dashed lines indicate negative regulation (silencing) toward the directed splice site. Hash marks indicate sequences not shown. Numbering along the top is HIV-1_{NL4-3} RNA numbering.

B. Results

1. The effect of the HIV-1 3'ss SA1 RNA stem loop structure on viral splicing efficiency

The stem-loop structure at the first 3'ss SA1 (SLSA1) has five conserved base-pairing interactions in the full-length genomic RNAs of both HIV-1_{NL4-3} and SIVmac239 (172) (Pollom *et al*., submitted), suggesting this structure plays an important role in viral replication to provide selective pressure for its maintenance. Although genomic RNA is packaged and dimerized, these full-length RNA structures are presumably good models for the unspliced RNA transcript for *gag* and *gag-pro-pol* mRNA. A small difference of local pairing partners in the 5' untranslated region has been described between packaged/dimerized and mRNA forms (100). We tested the importance of SLSA1 by making mutations that disrupt the structure and then observed the effect of these mutations on viral fitness and mRNA splicing patterns. The following considerations were taken into account when designing the mutations to SLSA1: *i*) we were careful not to disrupt any of the described ESE elements within the region downstream of the SA1 site (48, 76), *ii*) neither of the nucleotide mutations affected the *gag-pro-pol* coding sequence, *iii*) both of the mutations were to alternative codons that were also found nearby in the sequence so as not to require any rare or non-viral codon usage. The resulting mutant sequence (SLSA1m) has two single nucleotide substitutions that disrupt the pairing interactions at SLSA1 (Figure 3.2).

To test the relative fitness of the mutant virus HIV-1_{SLSA1m} to the wild-type HIV-1_{NL4-3}, we infected CEMx174 cells with both viruses in a coculture assay to test for relative fitness (171). The change in replication of both viruses was quantified by heteroduplex tracking assay (HTA) analysis, which separates the two viruses based on

sequence due to mismatched interactions with a heterologous probe (145) (Figure 3.2). We observe a change in the composition of virus within the culture by the third day post-infection when the wild-type HIV-1_{NL4-3} starts to become more prevalent than the mutant HIV-1_{SLSA1m}, indicating a greater ability of HIV-1_{NL4-3} to replicate in the cell culture than HIV-1_{SLSA1m} (Figure 3.2). However, this viral fitness difference is modest, with HIV-1_{SLSA1m} continuing to replicate fairly well, but not better than, HIV-1_{NL4-3} in cells.

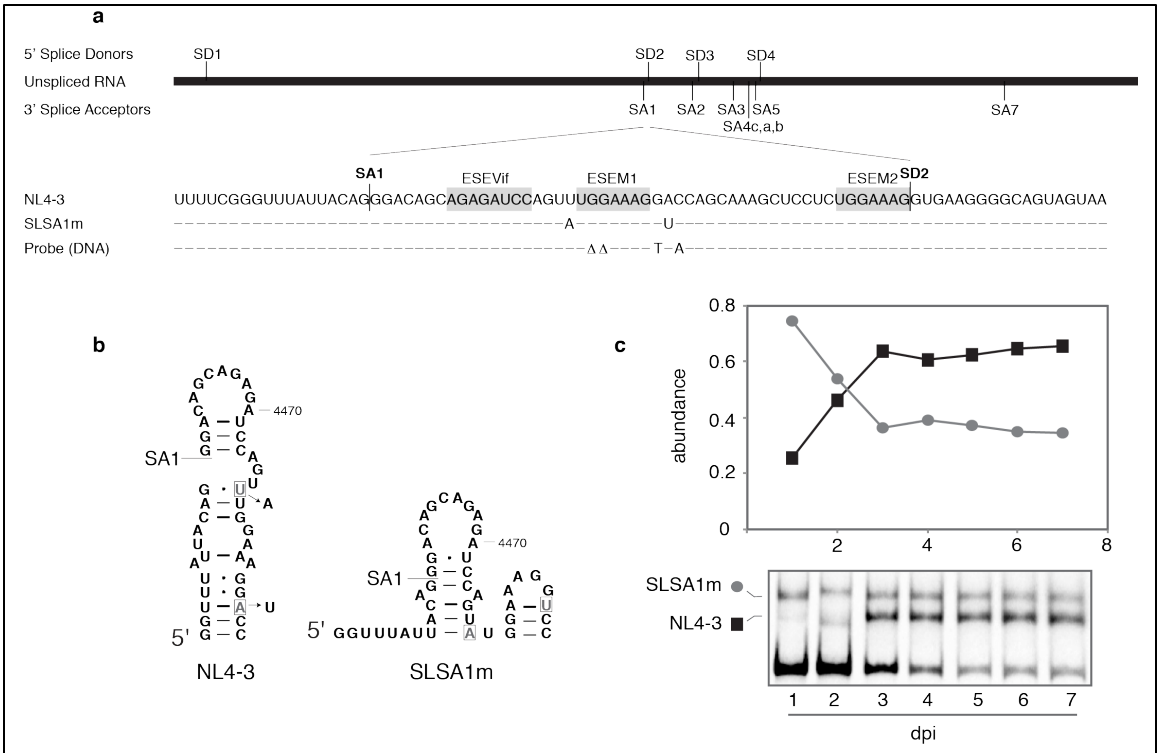


Figure 3.2: Mutations to the SLSA1 stem and coculture analysis of viral fitness. (a) Schematic of HIV-1 pre-mRNA showing 5'ss donors (SD1-4) and 3'ss acceptors (SA1-7) (above). Sequence around SA1 is shown with ESE regions highlighted. Single nucleotide mutations that disrupt the SLSA1 stem, producing SLSA1m, are indicated (A) and (U) while the sequence that remains the same is indicated by dashed lines. The sequence of the probe is indicated, where mutations and deletions are labeled (T), (A) and Δ, respectively (below). (b) Structures for the SLSA1 stem in HIV-1_{NL4-3} (left) and the SLSA1m stem in HIV-1_{SLSA1m} (right) as predicted using the *RNAStructure* (146) folding algorithm. The 3'ss SA1 is labeled and mutated nucleotides are boxed. (c) Heteroduplex tracking assay analysis of coculture with HIV-1_{NL4-3} (black squares) and HIV-1_{SLSA1} (gray circles). The cDNA products are separated based on sequence. Graph shows percent abundance of each virus at 1-day time points.

To explore possible causes of the slight decrease in viral replication between HIV-1_{SLSA1m} compared to HIV-1_{NL4-3}, we passaged both viruses in separate cell cultures for several days and examined the splicing profile of each viral mRNA pool. The mRNA was amplified with a forward primer that encompassed a unique NarI site at the 5' end and a reverse primer that was placed either after the 5'ss SD1 or the 3'ss SA7 to include either the 4 kb class or 1.8 kb class of spliced mRNAs, respectively (Figure 3.3). Each pool of cDNA products were cleaved with NarI and radiolabeled with ^{33}P-αdCTP, allowing each cDNA to be labeled only once at the 5' end and thus give uniform intensity of radioactive signal. Separating these products on a polyacrylamide gel gave a banding profile for each of three HIV-1_{NL4-3} cultures and three HIV-1_{SLSA1m} cultures (Figure 3.3). The cultures infected with HIV-1_{SLSA1m} have mRNA band intensities that are increased (Figure 3.3, right-directed arrows) or decreased (Figure 3.3, left-directed arrows) compared to the cultures infected with wild-type HIV-1_{NL4-3}.

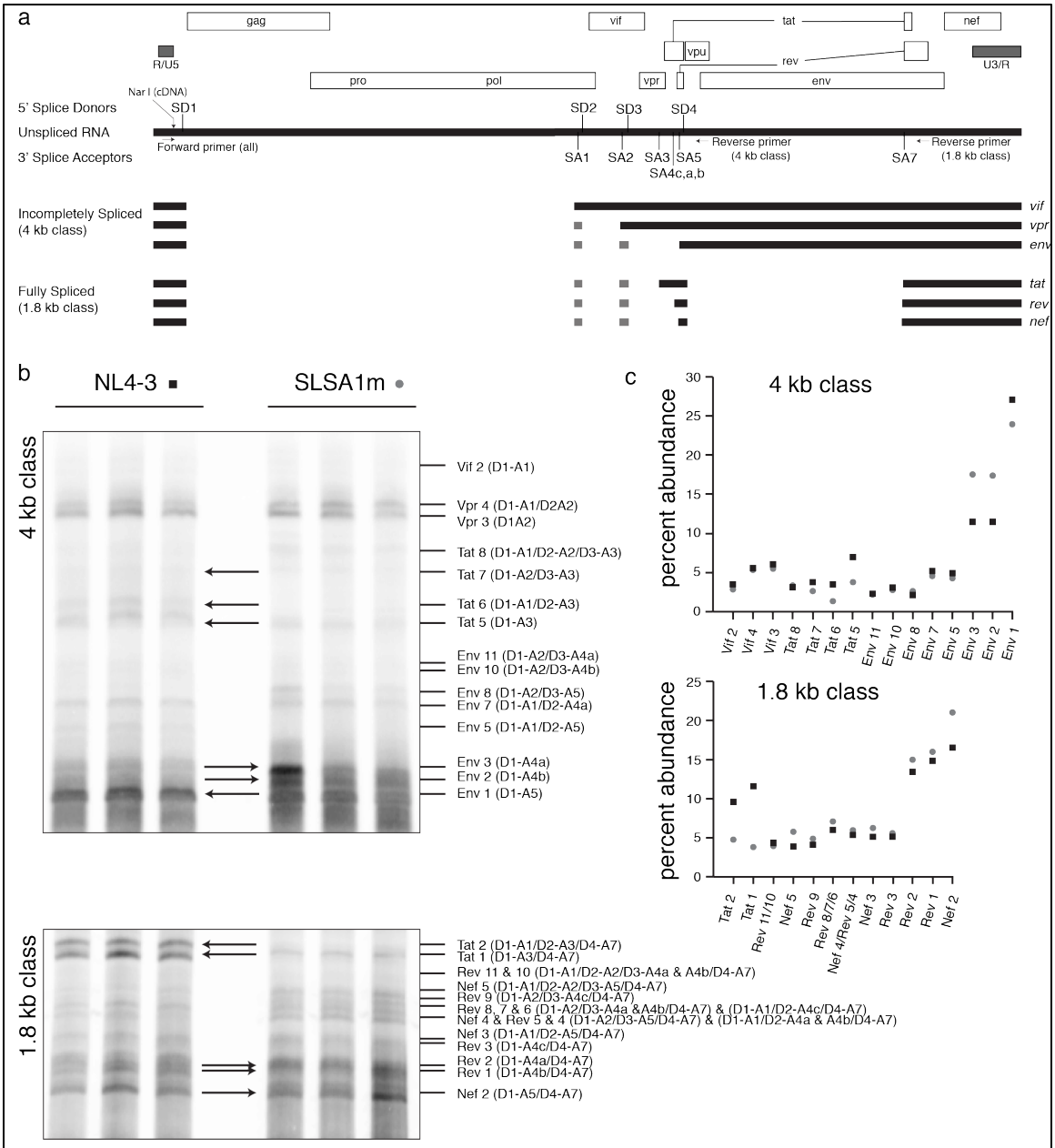


Figure 3.3: Profiles of HIV-1_{NL4-3} and HIV-1_{SLSA1m} transcripts. (a) Diagram displaying reading frames (open boxes) of the HIV-1 genome. Solid lines indicate different classes of mRNA including unspliced, 4 kb, and 1.8 kb, with their corresponding genes labeled on the right. Gray boxes represent exons 2 (between SA1 to SD2) and 3 (between SA2 and SD3). Splice donors (SD1-4) and acceptors (SA1-7) are labeled on the top of the unspliced length of RNA The sites of NarI cleavage and primer-binding for the forward and reverse primers used to create the splicing profile are shown on the unspliced RNA. (b) Splicing profiles for three separate cultures of HIV-1_{NL4-3} (black squares) and HIV-1_{SLSA1m} (gray circles) from primers that amplified the 4kb class (top) or 1.8 kb class (bottom) of mRNAs are shown. The cDNA was separated on a urea-containing polyacrylamide gel. The mRNA lengths are labeled according to common nomenclature (142). Arrows point to increased average abundance of bands. (c) Graphs of percent abundance of each variant of mRNA in either HIV-1_{NL4-3} (black squares) and HIV-1_{SLSA1m} (gray circles) that were visualized for the 4 kb class (above) and 1.8 kb class (below).

The splicing events that are augmented between the two viruses follow a pattern of fewer longer transcripts and, in general increased shorter transcripts in the HIV-1_{SLSA1m} cultures than in the HIV-1_{NL4-3} cultures (with the exception of the Env 1 variant). Longer transcripts include those which utilize 3'ss SA1, SA2, and SA3 and the shorter ones skip exons 2 and 3 and splice directly from 5'ss SD1 to SA4c/a/b or SA5. Specifically, the abundance of *tat* mRNA in the HIV-1_{SLSA1m} cultures decreases greatly compared to the amount of *tat* mRNA in HIV-1_{NL4-3} cultures. However, because SD1 can be spliced to either SA1 or SA2 followed by a splicing event between downstream donors and SA3 and still produce *tat* mRNA, it is not clear that the one specific splice acceptor or donor is preferentially used to any exact amount than another. What can be concluded, however, is that this alternate splicing profile produced by disruption of HIV-1_{SLSA1} RNA structure leads to a shift to the balance of necessary spliced mRNA products, which is a possible reason that the fitness of the mutant virus decreases compared to that of wild-type.

2. Similar features of RNA secondary structure in spliced mRNA variants

Previously, investigations toward understanding how RNA structure affects HIV-1 splicing regulation have been performed with small lengths of RNA, but as evidenced by our SLSA1 analysis, many features of secondary structure can be identified when the full-length of the RNA is analyzed. Therefore, we sought to map the full-length spliced mRNA structures to reveal similarities and differences between them. We made *in vitro* RNA transcripts of the reported most abundant spliced variants (142). Each variant transcript was constructed to reflect a splicing event between the major donor, 5'ss SD1,

and the 3'ss directly upstream of the given protein initiation codon (with the exception of the *rev* construct, which utilized the 3'ss SA4a, which is used slightly more frequently than the downstream 3'ss SA4b (142)). Although in cells, the multiply spliced RNA variants could potentially contain exons formed by splicing with SA1, SA2, SD2 and/or SD3 (Figure 3.3, gray exons), the main variant of each excludes these additional exons. The constructs of each multiply spliced variant also reflected a splicing event from 5'ss SD4 to 3'ss SD7, eliminating the large intron containing the Rev responsive element (RRE) and placing these transcripts in the 1.8 kb class of mRNA variants. The singly-spliced variants did not include this splicing event, giving them a longer sequence and placing them in the 4 kb class of mRNA variants (142). Throughout this section, we compare the SHAPE-derived structures of the HIV-1 mRNA variants to the SHAPE-derived HIV-1 full-length genomic RNA (172) with box-plot normalized data and folding parameters m = 1.9 b = -0.7 (Pollom *et al*., submitted). The mRNA transcripts included a tail of 20 adenosine nucleotides, and this tail was also used to purify the RNAs from the *in vitro* transcription reaction via an oligo dT affinity column. The final step in the purification method requires heating the RNA to 75°C. The RNA was allowed to refold at 37°C with no further denaturating chemicals or heat before the addition of 1M7.

Certain portions of the structure are maintained both between the different mRNAs and with the full length RNA. Much of the structure 3' of SA7 in the 1.8 kb class and 3' of SD4 of the 4 kb class is very similar. This includes an almost identical structure from nucleotide 8578 to the 3' end, which is likely due to the shared sequence in all of the variants after the common splicing event between SD4 and SA7 that occurs in the 1.8 kb class of mRNA. Similarly, the functional structures in the 5' region of all of these

mRNAs remain intact. Though the spliced variants lose major lengths of intronic sequences right after SD1, they keep some of the regulatory structures that are found in the 5' region of full-length genomic RNA. These regulatory regions contain stem-loops at the transacting responsive site (TAR), 5' poly (A) site, primer-binding site (PBS), and the dimer initiation site (DIS) and splice donor site (SD) hairpins (10, 36, 62, 131, 172, 173). The pairing interactions at the TAR, 5' poly (A), PBS, and DIS hairpins are mostly consistent with those seen in the full-length structure, with the exception of the DIS stem which forms long-distance pairing interactions with downstream sequence in *nef* mRNA (Figure 3.4a). The main changes that occur in the 5' region are to the loop of the PBS which, without a tRNA primer bound, forms different local interactions and to the SD stem, which loses half of its sequence after splicing. Slight changes are seen in the 5' poly(A) stem within the loop regions. The pairing remains constant throughout most of the 5' poly(A) stem, but the nucleotides that interact in a pseudoknot structure with the sequence in the coding region (134) are unreactive to the SHAPE reagent in the full-length RNA but more reactive in the spliced mRNA, with the exception of *vpu*/*env* mRNA whose sequence in this region remains unreactive. The interacting sequence was in the Gag-coding sequence, which is on the excised intron in spliced transcripts. Its removal leaves the 5' poly(A) loop unpaired and reactive. The loop region of the DIS stem maintains low reactivity even in the spliced structures, which implies that these transcripts are interacting with each other and is consistent with the observation by Sinck *et al* that the HIV-1 mRNA forms homo- and heterodimers with each other and with the genomic RNA *in vitro* (154).

RNA structure has been implied in splicing regulation with the function of bringing donors and acceptors closer together (24). We observe a possible example of this in the full-length genomic RNA and *vif* transcript with a large structured region between SD2 and SA2. Although the exact pairing interactions are not kept constant between the two configurations, the pairs at the base of the large structures are the same (Figure 3.4c, gray box). We hypothesize that this could be a previously undescribed splicing regulatory mechanism for HIV-1 and suggest that future studies, which disrupt the maintained pairing, would help confirm this supposition.

Watts *et al*. described a conserved stem at the signal peptide (SP) coding region of the Env protein and hypothesized a functional role for this region in stalling the translating ribosome, thus facilitating interaction of the Env SP to the signal recognition particle (SRP) and directing the translation complex to the ribosome (158, 172). Because this event would occur during translation from the *vpu*/*env* mRNA template, this hypothesis would only prevail if this same structure were present in the spliced transcript for the *vpu*/*env* gene. Indeed, the structure of *vpu*/*env* mRNA maintains this stem around the SP site that was previously described. The presence of this stem in the *vpu*/*env* mRNA gives credibility to the stalling ribosome hypothesis and would be a suitable template for further experiments to confirm this hypothesis involving footprinting analysis to detect paused ribosomes on the message.

Although many differences occur between the genomic and messenger RNA structures, many of these are simply due to the overall metastable state of the RNA (Pollom *et al.*, submitted). Notable differences seen between the genomic and transcript RNA mainly occur around sites where splicing has taken place. All of the spliced

messages depend on usage of 5'ss SD1, which occurs in the genomic RNA as the loop of a stem. When this donor is utilized, the structure at that stem is disrupted. Similarly, the splice acceptors SA1, SA3, and SA4a are contained in individual stems (SA2 and SA5 are in single stranded regions between distinctive structures) (Figure 3.4 – genomic). The result of splicing these sites is a change in the pairing interactions but a maintenance of base-pair interactions in the acceptor sites that were previously in stems and single-stranded regions in the acceptor sites that were previously single-stranded (Figure 3.4). This results in maintenance of many local interactions including structures that constrain or expose SRE sequences. This idea is particularly visible in the structure after the SD4/SA7 junction in the 1.8 kb class of mRNAs. The pairing or single-stranded characteristics of the regulatory sequences in the unspliced or incompletely spliced RNAs are maintained in this region regardless of the SD4-SA7 spicing event (Figure 3.4g).

Primer Name	Primer Sequence
HIV16	CCCTGACCCAAATGCCAGTCTC
HIV17	GCTCCCTCTGTGGCCCTTGGTC
HIV18	ATGAGCTCTTCGTCGCTGTCTCC
HIV19	CCCCATTTCCACCCCCATCTCC
HIV20	GTGGGGTTAATTTTACACATGG
HIV21	GAATCGCAAAACCAGCCGGGGC
HIV22	CATTTTGCTCTACTAATGTTAC
HIV23	CATCTCTTGTTAATAGCAGCCC
HIV24	TCTGGCCTGTACCGTCAGCGTC
HIV25	CTCTGTCCCACTCCATCCAGGTC
HIV26	CCTACCAAGCCTCCTACTATCA
HIV27	CTATTCCTTCGGGCCTGTCGG
HIV28	GCAAAATCCTTTCCAAGCCCTG
HIV29	GTAGCCTTGTGTGTGGTAGATCC
HIV30	GTACAGGCAAAAAGCAGCTGC
HIV31	TTTTTTTTTTTTTTTTTTTTTGAAG

Table 3.1: Sequences of primers used for SHAPE analysis of SIVmac239. Primers were designed and used for SHAPE analysis of the 3' end of the HIV-1_{NL4-3} genome (172) with the exception of HIV18, which was modified to accommodate splicing events.

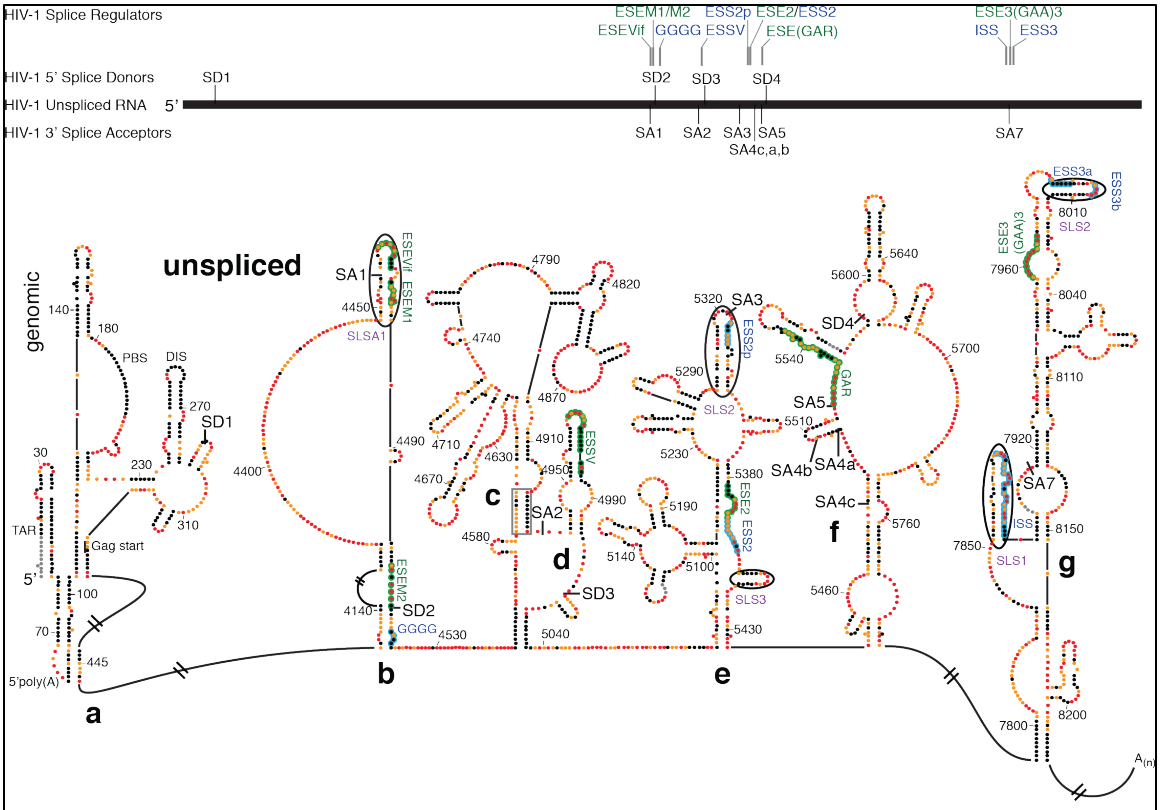
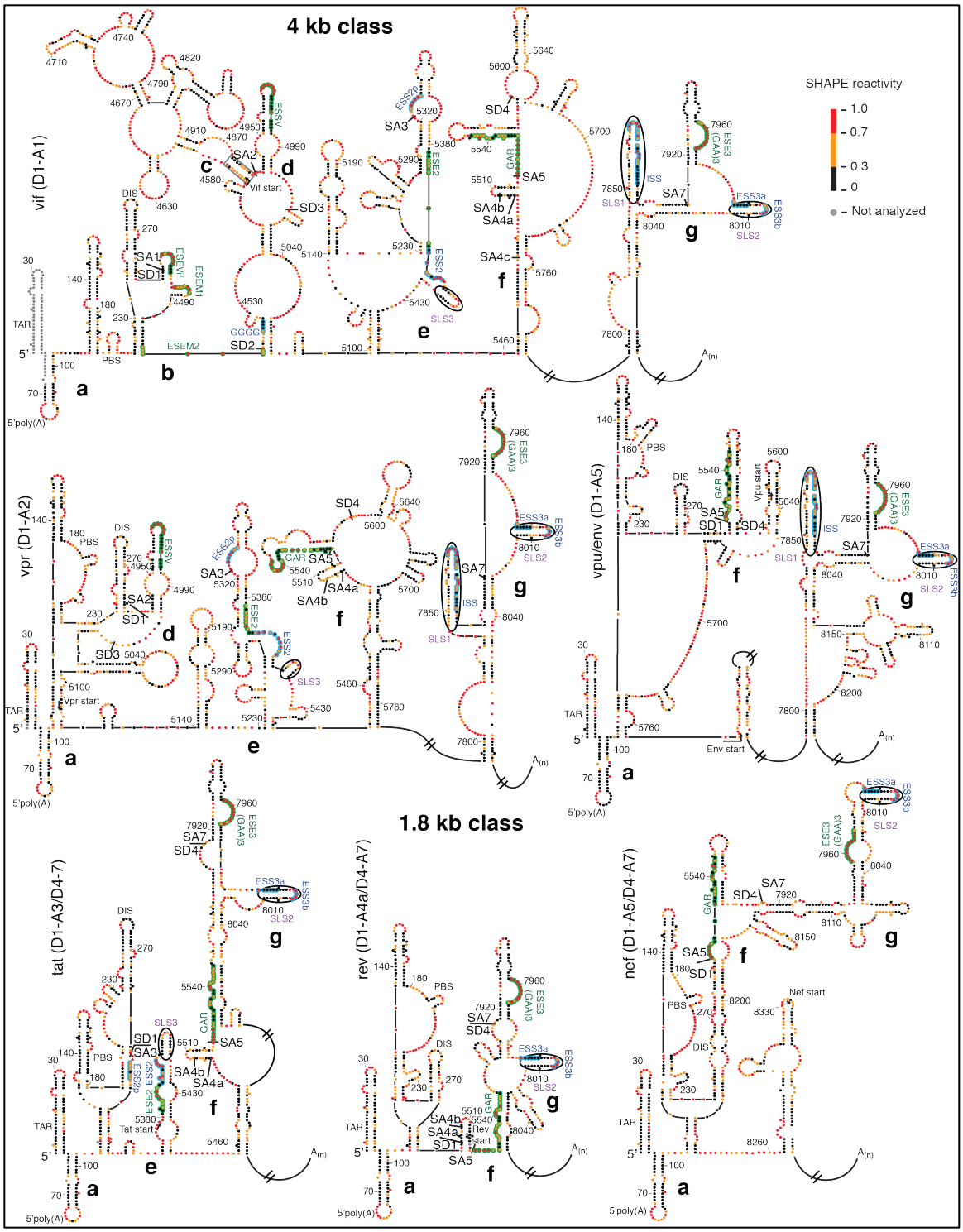


Figure 3.4: Structures of the most abundant variants of HIV-1 spliced messenger RNAs. Structures for the unspliced genomic, 4 kb class (*vif, vpr,* and *vpu/env*) mRNA, and 1.8 kb class (*tat, rev,* and *nef*) mRNA are shown. Nucleotides are displayed as dots and colored based on their SHAPE reactivity (scale shown) (SHAPE Primer sequences in Table 3.1). RNA is numbered according to full-length NL4-3 RNA sequence numbering. Hashed lines indicate regions of sequence that are not included in the figure. SLSA1 and previously described stem-loop structures are circled and labeled (purple). Regulatory structures in the 5' UTR are labeled. Start codons for the given transcript are labeled (TAR: trans-acting response element, poly(A): polyadenylation signal stem, PBS: primer binding site, DIS: dimerization initiation site). Splice donors, acceptors, splicing enhancers (green), and splicing silencers (blue) that appear in each structure are labeled (as depicted in top graphic). (a) 5' untranslated region, including SD1; (b) splice sites SA1 and SD2 with regulatory regions ESEVif, ESEM1 and ESEM2; (c) structure between SD2 and SA2 maintained between genomic RNA and *vif* mRNA, gray box indicates maintained base-pairs at the base of the structure; (d) splice sites SA2 and SD3 with regulatory sequence ESSV; (e) splice site SA3 with regulatory regions ESS2p, ESE2, and ESS2; (f) splice sites SA4c/a/b, SA5, and SD4 with regulatory region GAR; (g) splice site SA7 with regulatory regions ISS, ESE3(GAA)3, and ESS3(a/b). Figure continues on next page.


4 kb class
1.8 kb class
SHAPE reactivity
1.0
0.7
0.3
0
Not analyzed
vif (D1-A1)
vpr (D1-A2)
vpu/env (D1-A5)
tat (D1-A3/D4-7)
rev (D1-A4a/D4-A7)
nef (D1-A5/D4-A7)
TAR
PBS
DIS
SA1
SD1
SA2
SD2
SA3
SD3
SA4a
SA4b
SA4c
SD4
SA5
SA7
SD3
ESEM1
ESEM2
ESEVi
ESEV
ESE2
ESE3 (GAA)3
ESS2a
ESS2
ESS3
ESS3a
ESS3b
SLS1
SLS2
SLS3
ISS
GAR
GGGG
Vif start
Vpr start
Vpu start
Env start
Tat start
Rev start
Nef start
5'poly(A)
A(n)
a
b
c
d
e
f
g

3. Analysis of *cis* regulatory structures in spliced mRNA

Regulators of splicing must function in both the full-length transcript and in the incompletely spliced transcripts. Thus we compared the structural context of these elements in the unspliced and partially spliced RNAs and also observed changes to these regions in the completely spliced 1.8 kb class of RNAs. A summary of whether the majority of the nucleotides in the regulatory sequences form base paired or single stranded structures in these RNAs is given in Table 3.2. Overall, regulators that govern SA7 and SA2 along with ESEVif and ESEM1 which enhance splicing at SA1 and SD2 remain consistently either paired or single stranded in all the transcripts, but the SREs that regulate SA3 along with ESEM2 and the GGGG silencer element have altered structural conformation between genomic and messenger RNA (Table 3.2). The changed structures indicate a shift in the ability for the cellular proteins to recognize many of these regions after the initial SD1-3'ss event takes place. A consideration of each site follows.

	genomic	vif	vpr	vpu/ env	tat	rev	nef
ESEVif	ss	ss	-	-	-	-	-
ESEM1	ss	ss	-	-	-	-	-
ESEM2	bp	bp	-	-	-	-	-
GGGG	ss	bp*	-	-	-	-	-
ESSV	bp	bp	bp	-	-	-	-
ESS2p	bp	ss*	ss*	-	-	-	-
ESE2	ss	bp*	bp*	-	ss	-	-
ESS2	ss	bp*	bp*	-	bp*	-	-
ESE(GAR)	bp	bp	bp	bp	bp	bp	bp
ISS	bp	bp	bp	bp	-	-	-
ESE3(GAA)3	ss	ss	ss	ss	ss	ss	ss
ESS3a	bp	bp	bp	bp	bp	bp	bp
ESS3b	ss	ss	ss	ss	ss	ss	ss

Table 3.2: Comparison of SRE sequences in genomic and messenger RNA structures. SRE sequences (rows) are labeled as indicated in the text. RNA variants (columns) are labeled. Abbreviations ss and bp indicate whether most of the sequence is single stranded or base paired, respectively. Changes in base-pairing or single-strandedness compared to the genomic structure are indicated with an asterisk (*). Gray rows signify enhancer elements and white rows signify silencer elements.

Splicing events at 3'ss SA1 and 5'ss SD2 are influenced by multiple sequences in the exon between these sites which share similar structural motifs in both the genomic RNA and *vif* mRNA. Three enhancer elements, ESEVif, ESEM1, and ESEM2, promote splicing at SD2, and ESEVif also promotes SA1 recognition (48, 76). Even though the intron between SD1 and SA1 is excluded in the *vif* mRNA, dramatically changing the sequence directly before SA1, a common theme is still maintained between the enhancer elements after SA1: each enhancer sequence is mostly located within a single-stranded loop except ESEM2 of genomic RNA (Figure 3.4b). ESEVif is found in the loop region of SLSA1 in the genomic structure and the loop region of an analogous stem in *vif* mRNA. ESEM1 is located in a bulge of SLSA1 in the full-length structure and in the loop of a small stem in the *vif* mRNA. ESEM2 has the common theme of being incorporated between two stems in both structures, though it is contained within more pairs in *vif* mRNA. The role of these enhancer sequences is to bind either SR proteins SRp75 (ESEVif) (48) or SF2/ASF (ESEM1 and ESEM2) (76), so their single-stranded sequences might improve their exposure to these splice-promoting proteins. After the first splicing event between SD1 and SA1, the GGGG silencer is more secluded in base pairs. This could disallow the silencer element to be recognized, upregulating use of SD2 when SA1 is used. It is of note that ESEVif and ESEM1 are both incorporated into SLSA1, both of which occur directly upstream of the conserved pairing interaction between HIV-1_{NL4-3} and SIVmac239 that was described (Pollom *et al*., submitted). This pairing interaction could potentially perform the function of maintaining these sequences in their respective structures.

A sequence downstream of the 3'ss SA2 has been previously reported to act as an exonic splicing silencer (ESSV) (13, 103). We mapped this 16 nt sequence onto the full-length genomic RNA and the *vif* and *vpr* mRNA structures to find an exactly identical stem-loop structure shared by all three forms of the RNA (Figure 3.4d). The local sequence in this region is not altered in *vpr* mRNA since it is far enough downstream of SA2, making any structural rearrangement less likely. The pairing interactions at this region have been described (149), but the effect this structure has on the regulatory sequence is yet unknown.

Expression of *tat* mRNA has been described as being regulated by ESS2p at a small stem-loop structure followed by ESE2 and ESS2 elements on a long irregular stem further downstream of 5'ss SA3 (3, 61, 71, 176), but we have found slightly divergent structures in our analysis. We examined the mRNA variants that would contain the sequences necessary to form these previously predicted structures: genomic, *vif*, *vpr*, and *tat* mRNA. The genomic RNA structure includes the short stem containing ESS2p. However, although both *vif* and *vpr* mRNA have structures similar to each other in this region, the long irregular stems formed by these singly spliced mRNAs are not the same as the one that has been previously described (Figure 3.4e). The ESE2 and ESS2 enhancer sites in the genomic RNA are both single-stranded, allowing for use of either, depending on levels of SR or hnRNP proteins in the cell. After splicing at SA1 or SA2, the silencers become single-stranded while the enhancers are secluded in pairing interactions (Figure 3.4e). This may allow for binding of hnRNP factors, but not the enhancing SR proteins, thus limiting the amount of *tat* transcript and subsequent Tat protein production.

The ESE GAR sequence that positively influences splicing at all of the surrounding splice sites, SA4c/a/b, SA5, and SD4 is mostly paired in all of the structures (Figure 3.4f). This sequence works on the acceptors before it and the 5'ss SD4. No structure has been attributed to regulation for this region, but we observe pairing interactions at many of the nucleotides involved in the enhancer sequence in all of the transcripts and the genomic RNA. Utilization of the upstream acceptors SA4c/a/b and SA5 does not differentially affect the use of SD4 given that any of these acceptors can be used without the splicing event occurring from SD4 to SA7 (142). Therefore, the regulation from the GAR sequence to these splice sites is not necessarily linked. Sequence could be secluded in this structure to minimize the efficiency of the enhancer toward any of its surrounding splice sites.

Each of the singly spliced mRNA structures as well as the full-length genomic RNA structure contain the same motif around the region of the 5'ss SA7, but slightly different interactions than what has been previously described. At this site, regulation elements termed ISS and ESS3 perform the function of silencing the splicing event between SD4 and SA7 by binding hnRNP A1 (5, 37, 111). Regulatory element ESE3/(GAA)3 both silences and enhances this splicing event (37, 111). Three stem-loop structures at this region have been previously mapped using chemical probing analysis (37), and recently one of the stem structures, SLS2, was solved by NMR (96). Each of these methods, however, has been performed using small isolated lengths of RNA exactly surrounding the site of interest. Our analysis reveals slightly different structures in this region due to the availability of interactions at other sequences that are outside the previously examined range (Figure 3.4g). The singly-spliced mRNA structures form very

similar interactions to those previously described, but the full-length genomic RNA makes slightly different interactions. All share the proposed structure at SLS1 around the ISS sites, and SLS3 which includes the ESS3 sites (Figure 3.4g). However, the singly-spliced mRNA forms a long stem structure at the ESE2 site which is similar but not identical to the SLS2 stem described (96) while the full-length genomic RNA and *nef* mRNA do not form a stem-loop at this site, but a long helix leading to SLS3 instead (Figure 3.4g). This helix does contain an important element, however: the bulge in the helix is the same as that containing the GAA repeats implicated in ESE2. The multiply-spliced variants also maintain both the SLS3 bulges as well as an intact SL3 with the ESS3 elements. This indicates that the pairing interactions of the shorter stems should be preserved for function, but the long stem structure may not be as important in regulation as the maintenance of a single-stranded loop at the ESE3/(GAA)3 site.

To further characterize how RNA structure influences regulation of pre-mRNA splicing in lentiviruses, we compared the structures of HIV-1 genomic and messenger RNA to the structure of SIVmac239 genomic RNA (Pollom *et al*., submitted). Most of the structures around the splice sites in SIVmac239 differed from those of HIV-1. Sequence analysis has shown that most of the SRE sequences in different clades of HIV-1 are strongly conserved (reviewed in (160)), so we examined the sequence and structure of SIVmac239 to find that only the SRE sequences and the pairing patterns of these sequences around SA1 are conserved. These discrepancies between SIVmac239 and HIV-1 could be a result of the different protein-coding regions in the viruses. SIVmac239 RNA contains the *vpx* gene (whose 3'ss directly upstream is SIV's second splice acceptor) and lacks the *vpu* gene. The viruses could be therefore differentially regulated

to account for these discrepancies. The two splicing events that are more similar between the viruses are the event that takes place from SD1 to SA1 to produce *vif* mRNA (as we have investigated above) and the event that excludes the RRE-containing intron. Structures in SIVmac239 at both of these regions not only have low median SHAPE reactivity, but are also high in guanosine content compared to adenosine content over a 75 nucleotide window (Pollom *et al*., submitted). Although the pairing partners are not conserved in the SIVmac239 3'ss SA8 (whose equivalent is SA7 in HIV-1), the strong pairing interactions indicated by SHAPE reactivity and G-content suggest that the structure around this splice site performs a splicing regulatory function similar to that of HIV-1.

C. Discussion

1. Altering the structure at SLSA1 has a moderate effect on viral replication and influences the splicing profile of HIV-1

The effect of mutating the stem-loop structure at 5'ss SA1, which we have termed SLSA1, is not a drastic decrease in viral replication. Instead, the mutation causes a moderate change in the fitness of the virus due to an alteration in the pattern of mRNA splicing. The pattern, though different from the wild-type splicing pattern, still produces the necessary messages to maintain replication but at disparate levels from that of wild-type HIV-1_{NL3-4}. This is perhaps because all of the splicing factors work in concert to effect viral splicing. Changing one of these factors might force the others to silence or enhance splicing at a different level to keep the balance needed for every spliced message to be transcribed. Furthermore, with only two mutations, the SLSA1 may maintain some of its stability, but disrupting the stem further by modulating more of the sequence would

also disrupt known splicing regulatory factors contained therein. The mutations that were made in this study would have only disrupted the stem without changing the amino acid sequence or using any rare codons, making it possible for us to conclude that the SLSA1 structure plays a role in regulating alternative splicing of HIV-1. The pairing partners that are conserved between HIV-1_{NL4-3} and SIVmac239 include the nucleotides involved in the polypyrimidine tract of SLSA1. We suggest that SLSA1 performs the function of secluding the polypyrimidine tract of SA1 from the spliceosomal factors that recognize this sequence.

2. Analysis of regulatory structures that are maintained in spliced mRNA

Despite excision of the intron after SD1 to the given splice acceptor in all of the mRNA products, the spliced mRNA variants keep the functional structures at the 5' UTR that are intrinsic in full-length genomic RNA. The TAR structure performs the same function of transcription initiation in genomic mRNA as it does in full-length RNA, but the functions of the PBS and DIS stems seem unnecessary in the spliced variants. The presence of these stems could simply be a remnant of their strength and stability, implying that these local structures are necessary to the virus and even large sequence omissions or changes do not have a deleterious effect on them.

Many of the elements implicated in HIV-1 mRNA splicing regulation have been described as sequences or small independent structures. However, mapping these sequences onto the SHAPE-derived mRNA structures and comparing how they change based on splicing events can give evidence of importance in different stages during splicing. Some of these structures have remained intact while others vary based on

changes in structure due to different local sequences. Furthermore, we observe structures that have been yet undefined at regulatory sequences ESEVif, ESEM1, and GAR.

Although the splicing events from SD1 to the given 3'ss and between SD4 to SA7 have already occurred in the fully spliced variants, they still maintain many of the splicing silencer and enhancer features of the incompletely spliced structures. This is peculiar due to the assumption that these structures would act similarly in all of the transcripts, limiting or enhancing splicing identically when the structures are the same. However, the presence of these structures strengthens the idea that HIV-1 RNA splicing is an act of balance. The silencing elements will potentially be impeded by enhancers, and vice versa. The regulatory structures we observe are typically maintained in small hairpins. Even though HIV-1 RNA folding has been hypothesized to occur post-transcriptionally (177), short-range structures would likely form first and become stable before long-range structures had a chance to interact both in cells and *in vitro*. Therefore, these small hairpins would only depend on the local sequence around them. Furthermore, these mRNA structures were analyzed in the absence of RNA binding proteins. Such splicing factors could act to strengthen or weaken the given structures, forcing the regulatory sequences into different conformations. The virus relies on the equilibrium of these forces to produce the adequate heterogeneous mixture of all the diverse mRNA variants. We conclude that many of the regulatory elements essential for this transcript diversity are impervious to the sequence alteration that follows viral mRNA splicing.

Structures at the regulatory sequences around SA7 are similar to what has been previously shown, but the exact pairing interactions are different. These previous studies used shorter lengths of RNA, which may not have been sufficient to resolve the other

local and long-range interactions that are occurring. Although the structure is rearranged, the described necessary loop and paired regions remain the same. These are the binding sites for hnRNP A1, particularly the loop region of ESE3(GAA)3, which is the initiation site for hnRNP A1 binding (37, 111). After this initial cellular protein binds, the surrounding structures may be effected and change to better resemble the previously published RNA structures (5, 37, 111).

The regulatory structure SLSA1 that we describe here is not the only regulator of splicing at the 5'ss SA1. The other sequences found around SA1 that bind to regulatory SR proteins include ESEVif, ESEM1, and ESEM2 (48, 76). We found all three of these sequences to be in partially single-stranded regions in both full-length genomic RNA and *vif* mRNA SHAPE-derived structures, even though the exact pairing partners for the surrounding base-paired interactions are different due to the altered sequence brought about by SD1-SA1 splicing. This suggests that these regions are not only available for protein binding, but their maintenance in both RNAs implies that they are used in regulating splicing events regardless of whether 3'ss SA1 is used or not. This corresponds to the observation that ESEM1 and ESEM2 have been implicated in regulation of 3'ss SD2 usage, but are not responsible for diminishing the amounts of singly-spliced *vif* mRNA (76). Therefore, having the same structural motif as in full-length genomic RNA implies that their function is the same for both. However, the change of the silencer GGGG sequence from being single stranded in the genomic structure to being paired in the *vif* mRNA structure indicates a possible lowered recognition of this region, potentially decreasing its silencing ability toward SD2 after the first splicing event.

Given that mutations within the nine nucleotides upstream of the necessary ESSV sequence did not affect the silencing function of that region (103), even though these account for the 5' side of the observed stem, perhaps the sequence is more important than the structure for exonic silencing. However, the maintenance of the exact pairing partners in this stem within the genomic RNA and in *vif* and *vpr* transcripts strongly suggests a function for this structure. Perhaps it contains some other splicing enhancer in its vicinity that the *vpr* transcript requires to cooperatively counteract the silencing function of the sequence.

The described stem structures around SA3 contain silencing elements ESS2p in SLS2 and ESS2 in SLS3 that bind hnRNP H and hnRNP A/E, respectively, and function to counteract the enhancing ability of ESE2 in SLS3 which binds SC35 (3, 61, 71, 176). Although none of the spliced messages maintain SLS2, the longer SLS3 structure is present in *tat* mRNA while *vif* and *vpr* only share the small stem loop that is formed at the end. The *vif* and *vpr* mRNA contain a stem-loop structure upstream of ESS2 which forces ESE2 into another stem while ESS2p is contained within a single-stranded region. We propose that this structural rearrangement is another method of regulating the production of Tat. If either of the first two splice acceptors are used, the silencers around SA3 become more accessible to regulatory proteins while the enhancer sequence is secluded, disallowing further splicing from SD2 or SD3 to SA3.

The newly discovered structure at the first 3'ss SA1, termed SLSA1, seems to play a role in HIV-1 pre-mRNA splicing regulation. This is just a single part in the complex coordinated splicing regulation scheme that is necessary to produce the adequate levels and variety of the HIV-1 transcriptome. The full-length mRNA structural models

presented here give a wide view of the impact of structural maintenance or change based on intron excision and provide evidence that these structural motifs are necessary for splicing regulation and essential for proper HIV-1 replication.

D. Materials and Methods

1. Cell lines

CEMx174 and 293T cell lines were obtained from the National Institutes of Health ADS Research and Reference Reagent Program. CEMx174 cells were sustained in RPMI 1640 medium with 10% fetal calf serum and penicillin-streptomycin. 293T cells were maintained in Dulbecco's modified Eagle medium with 10% fetal calf serum and penicillin-streptomycin.

2. Site-directed mutagenesis

The viral plasmid pNL4-3 was acquired from the National Institutes of Health ADS Research and Reference Reagent Program. For site-directed mutagenesis of pNL4-3, fragments digested with PflMI and AgeI (New England Biolabs) were inserted into vector pT7Blue (Novagen). Mutagenesis primers 5'-GAGATCCAGTATGGAAAGGTCCAGCAAAGCTCCTC-3' and 5'-GCTTTGCTGGACCTTTCCATACTGGATCTCTGCTG-3' were used in accordance with the previously described mutagenesis protocol (89). The resulting plasmid and pNL4-3 were then digested with PflMI and AgeI (New England Biolabs) and ligated with T4 DNA Ligase (New England Biolabs) to create the plasmid pSLSA1m, which was sequenced to confirm the presence of the given mutation.

3. Virus production

A total of 2 μg of pSLSA1m mutant or wild-type viral plasmid pNL4-3 and was used to produce mutant and wild-type viruses by transfection into $3x10^5$ 293T cells in a volume of 2 ml DMEM following the FuGENE (Promega) protocol. After 48 hrs, supernatant from the cells was centrifuged, transferred to 1 ml aliquots, then stored at -80 °C. One aliquot per virus was used in a viral infectivity assay (82) to determine infectious units per ml of supernatant.

4. Isolation of viral mRNA from cells.

To obtain viral mRNA, 5 x 10^5 cells were infected with 0.3 ml virus (either wild-type or mutant) supernatant in a volume of 0.5 ml for 2 hrs at 37 °C before being brought to a final volume of 10 ml and incubated at 37 °C for four days. Cells were centrifuged and supernatant was removed. The cell pellet was homogenized through a QiaShredder column (Qiagen) and total mRNA was purified by the RNeasy Mini Prep kit (Qiagen) according to the manufacturer's protocol.

5. Viral mRNA profile

Using viral mRNA isolated from three flasks of cells infected with HIV-1_{NL4-3} and three flasks of cells infected with HIV-1_{SLSA1m}, we digested each sample with RQ1 DNase (Promega) for 2 hrs at 37 °C and purified them again using the same RNeasy Mini Prep kit (Qiagen). We then performed One-Step RT-PCR (Qiagen) following the manufacturer's protocol and using primers 5'-

AGTCAGTGTGGAAAATCTCTAGCAGTGG-3' and either 5'-CCGCAGATCGTCCCAGATAAG-3' (1.8 kb class) or 5'-CTATGATTACTATGGACCACAC-3' (4kb class) in a volume of 25 μl. Each was then digested at a unique restriction site with NarI (New England Biolabs) for 2 hr at 37 °C and labeled with 0.78 μCi ^{33}P-αdCTP (Perkin-Elmer) using Klenow fragment (New England Biolabs). Each was then purified through a PCR Purification column (Qiagen) and eluted with 30 μl elution buffer (Qiagen). A sample of 10 μl of each was mixed with 10 μl 90% formamide in 1xTBE and denatured by boiling for 2 min. Samples were run on a 7M Urea gel with 6% polyacrylamide.

6. Virus coculture

The HIV-1_{SLSA1m} mutant virus and HIV-1_{NL4-3} wild-type virus dually infected CEMx174 cells in a coculture-growth competition assay as described in (171) to measure relative fitness. Briefly, the two viruses were used to infect 3 x 10^6 CEMx174 cells in a ratio of 3:1 with 600 infectious units of the mutant virus and 200 infectious units of the wild-type. Infected cells, in a volume of 0.5 ml, were incubated at 37 °C before being washed with 1 ml 1x PBS, centrifuged, resuspended in 0.5 ml 1x trypsin (Sigma), incubated at 37 °C for 5 min, and centrifuged again. The cell-virus pellet was then resuspended in 10 ml medium. Virus supernatant samples were taken every day and the cells were resuspended in fresh medium. RNA extraction and PCR amplification was done as described (145).

7. HTA

Heteroduplex tracking assay (HTA) was performed to assess the relative ratio of HIV-1_{SLSA1m} mutant virus to HIV-1_{NL4-3} wild-type as described (145) with a modification of being labeled with ^{33}P-αdCTP (Perkin-Elmer) after cleavage with SpeI (New England Biolabs). The probe was designed to separate the two virus cDNA amplicons based on sequence at the site of mutation, and the sequence difference is shown in Figure 3.2a. Hybridized heteroduplex products were run on a 6% polyacrylamide gel.

8. Transcription template plasmid construction

To create a template for transcription, we amplified the 5'R sequence until the 5'ss SD1 by PCR of plasmid pNL4-3 using oligonucleotides 5'-AATAGCATGCGGTCTCTCTGGTTAGACCAGATCTGAGCC-3' and 5'-AATAATCTAGACTTTCAAGTCCCTGTTCGGGCGCCACTGCT-3' and the sequence from 3'ss SA7 to the 3'Repeat region including a 20-adenosine tail and AatII cleavage site by PCR amplification of plasmid pNL4-3 with oligonucleotides 5'-AATATCTAGAGTGCAGGGGAAAGAATAGTAGACATAATAG-3' and 5'-GACGTCTTTTTTTTTTTTTTTTTTTGAAGCACTCAAGGCAAGCTTTATTGAGGC-3'. The cDNA products were ligated separately into pT7Blue (Novagen) using T4 Ligase (New England Biolabs) to create pExVec. Subsequently, cDNA products for mRNA variants were produced by RT-PCR amplifying viral mRNA previously isolated from cells with the following primers: 5'-GAAAATCTCTAGCAGTGGCGCCCGAACAGG-3' and either 5'-CTCTTTTTCCTCCATTCTATGGAGACTCC-3' (SD1-SA1), 5'-

CGAGTAACGCCTATTCTGCTATGTCGAC-3' (SD1-SA2), 5'-CCAAATTGTTCTCTTAATTTGCTAGCTATC-3' (SD1-SA5), or 5'-GATCGTCCCAGATAAGTGCTAAGGATCC-3' (SD1-SA3/SA4a/SA5 and SD4-SA7) using the One-Step RT-PCR Kit (Qiagen) following the manufacturer's protocol. The resulting cDNA products and pExVec were digested with restriction enzymes (New England Biolabs) NarI and PflMI (SD1-SA1), SalI (SD1-SA2), NheI (SA1-SD5), or BamHI (SD1-SA3/SA4a/SA5 and SD4-SA7), then ligated using T4 Ligase (New England Biolabs) and screened for content by sequencing analysis with primers 5'-ATGACCATGATTACGCCAAG-3' and 5'-GGTTTTCCCAGTCACGACG-3'. This created plasmids pSVif, pSVpr, pSEnv, pSTat, pSNef, and pSRev.

9. RNA *in vitro* transcription.

To create a blunt stop for transcription, 1500 ng of each plasmid (pSVif, pSVpr, pSEnv, pSTat, pSNef, and pSRev) was linearized with one unit of restriction enzyme AatII (New England Biolabs) followed by end repair using *DNATerminator* (Lucigen) following the manufacturer's protocol. RNA was transcribed from 500 ng digested plasmid template using MegaScript® T7 (Ambion) following the manufacturer's protocol at 37 °C for 3 hrs. RNA was then DNase treated with 2 μl DNase Turbo (Ambion) at 37 °C for 15 mins.

10. RNA purification.

Transcribed RNA was purified using the GenElute™ mRNA Miniprep Kit (Sigma Aldrich) following manufacturer's protocol, which uses oligo(dT) beads to bind to and

purify RNA containing a poly(A) tail followed by heating and elution and precipitation with ethanol. Purified RNA yield was typically 25-50 ng/μl.

11. SHAPE analysis of RNA

RNA was modified as described in Chapter 2. Briefly, the purified RNA was precipitated and the ethanol was removed. Each pellet contained from 10 to 40 pmol HIV-1_{NL4-3} transcribed RNA of the given variant and was resuspended in 10 μl of 50 mM HEPES (pH 8.0), 200 mM potassium acetate (pH 8.0), 3 mM $MgCl_2$ per 1 pmol RNA. The RNA-containing solution was then incubated for 10 min at 22 °C then 15 min at 37 °C. Per 1 pmol RNA, 1 μl aliquots of 45 mM 1M7 in dimethyl sulfoxide (DMSO)(122) or DMSO alone were pre-warmed (37 °C) for 30 sec before addition of half the RNA solution to each, followed by incubation for 5 min. Per 1 pmol RNA, 1 μl 50 mM EDTA (pH 8.0) was added then the RNA was recovered by precipitation with ethanol.

12. Primers

Primers were labeled exactly as in Chapter 2. A total of 16 primers were used (Table 3.1).

13. Primer extension

Primer extension was performed exactly as in Chapter 2. Plasmids used for primer extension were the same as were used for transcription templates.

14. Data processing

Data were converted using ShapeFinder software (40, 164, 173) exactly as in Chapter 2. Data were processed using the box-plot normalization method (J. Low, personal communication).

15. RNA secondary structure modeling

Each mRNA sequence was folded in its entirety with using the *RNAstructure* algorithm (114, 146) as described in Chapter 2. No pairing or forced single-strands were constricted in any of the models. We used folding parameters m = 1.9 and b = -0.7 (C.E. Hajdin, personal communication) to generate structures for all six the mRNA variants.

16. RNA structure display

The secondary structure models for all of the mRNA were arranged using xrna (http://rna.ucsc.edu/rnacenter/xrna).

CHAPTER IV

CONCLUSION

In this work, I have described two studies that use the SHAPE chemical probing method as a tool for understanding the conservation, maintenance, and function of RNA structure in different stages of the lentiviral replication cycle. Given the SHAPE-derived structures of SIVmac239 genomic RNA and HIV-1 messenger RNA, I was able to probe further into the relevance of individual pairing interactions based on evolutionary conservation of specific regions, formation of RNA structure based on base composition, and changes or maintenance of structure in certain sites regardless of or as a function of splicing events.

Determination of the secondary structure of genomic SIVmac239 RNA provided a second lentiviral RNA structure with which I was able to compare regions of the genome that contained low SHAPE reactivity, high guanosine content, and high pairing probability to those of the published HIV-1_{NL4-3} structure (172). The results showed conserved RNA structures at regions with known functions in the viruses. Stems at the 5' UTR, frameshift, and RRE regions have been well established and were very similar. These also are included in the few areas of the RNA where guanosine concentration is higher than that of adenosine. These G-rich regions are islands of conserved structure in an otherwise A-rich and non-conserved genome. Even at these RNA structures with

known function, however, the exact pairing interactions are not identical between the two viruses. In fact, only a small percentage of base-pairs are maintained between the relatively similar genomes. I conclude that the RNA genomes of SIVmac239 and HIV-1_{NL4-3} fold in a way that is rapidly changing over the course of evolution, and the regions that are necessary for function are kept stable by localized guanosines that strengthen pairing interactions.

Analysis of the structure of HIV-1 messenger RNA allowed identification of structures that are maintained in an entirely separate stage of the lentiviral replication cycle than genomic RNA. Functional assays that tested the effects of mutating the evolutionarily conserved stem at the first splice acceptor site (SA1) showed the importance of that stem (termed SLSA1) in the complex splicing regulation of HIV-1 mRNA. Disruption of this previously undescribed stem only had a mild effect on the replication capacity and splicing profile of the virus, illustrating that many other factors contribute to the production of spliced products. I further analyzed the role of RNA structure in splicing regulation by determining the secondary structures of *in vitro* transcribed variants of HIV-1 mRNA. Many of the structural motifs were slightly different from what had been previously described due to the ability of the SHAPE method to allow structural probing of longer RNA molecules. Furthermore, many of the structures that I describe are maintained between the fully spliced, incompletely spliced, and unspliced variants unless a change in pairing is necessary to preclude downstream splicing events. From these results, I conclude that the complex regulation of HIV-1 mRNA splicing is controlled by RNA structures that interact locally and most are stable enough to be impervious to drastic changes and deletions throughout the sequence,

however, the regulatory sequences that control use of SA3 have altered structures after usage of the first two splice acceptors which may allow the singly-spliced SA1 and SA2 transcripts to downregulate further splicing at SA3.

The final product of this work broadens our understanding of the role of RNA structure in lentiviral replication, but is not a conclusive analysis of the subject. Instead, it develops a starting point for further research and investigation. The purpose of these RNA structural analyses was to gain a better understanding of structural conservation, rearrangement, and maintenance for the sake of identifying structures that may be important for various aspects of viral replication. Identification of these structures has, in effect, opened the doors to many new and potentially critical experiments and analyses.

The possibility of a conserved pseudoknot between the SIVmac239 and HIV-1 5'UTR and Gag-coding sequence based on SHAPE reactivity led to experiments involving locked nucleic acids (LNAs), which seclude this pairing interaction on both sides of the 5' poly(A) pseudoknot. The results of these experiments, not described herein, were too weak to support the conclusion of a pseudoknot at this site in SIVmac239 as has been identified in HIV-1. The SHAPE data at the site of interaction is strong evidence toward a pseudoknot interaction; however, further investigation into this site is necessary to claim with confidence that this pairing interaction is definitively conserved between the two viruses. Mutations that change the sequence of either side of the predicted pseudoknot in SIVmac239 followed by a change in the SHAPE reactivity of the predicted pair would strengthen this conclusion. Furthermore, mutations that disrupt the 5' poly(A) pseudoknot in SIVmac239 or HIV-1 may allow for a functional analysis of this interaction which explores its effect on viral replication in cells.

An obvious investigation that is drawn from the conclusions made in our SIVmac239 structural analysis is a comparison of the guanosine and adenosine abundances in other viruses. Lentiviruses have an A-rich genome, which makes identification of isolated regions of Gs fairly clear to observe. Identifying clusters of guanosines within other viral genomes may point to potentially significant functionally conserved structures as in HIV-1 and SIVmac239. Conversely, finding areas where adenosines are rich in G-dominant genomes would indicate lack of structure with possible functional relevance.

The decision to determine the genomic RNA structure of SIVmac239 was due to many reasons including convenience and accessibility of obtaining a large amount of purified viral RNA from Dr. Rob Gorelick and coworkers, the long evolutionary distance between HIV-1_{NL4-3} and SIVmac239, and the characterization of many known structures of SIVmac239. This choice, however, was also predicated on the idea that many of the structural motifs between the two viruses would be well conserved. Since this is not the case, structural analysis of a more closely related virus to HIV-1_{NL4-3}, perhaps SIVcpz or even another HIV-1 subtype would give more clues into the conservation of RNA secondary structure between lentiviruses. Based on our findings, the difference in structural conservation between two distantly related lentiviruses is not surprising. However, a structural change between two lentiviruses whose sequences were more highly conserved would enhance the argument that structure evolves more rapidly than sequence. If many of the structures stayed intact, this would lead to further analysis of the functional relevance of these regions.

An analysis that uses SHAPE to probe structures of the regulatory sites after the mutations at SLSA1 altered the stem structure would be beneficial to analyzing the effect of mutations to this region. Obtaining enough mutated full-length genomic RNA would be difficult, but inserting the mutation into the *vif* transcript would allow us to observe how this mutation is affecting the structure, even after splicing. Disrupting the stem at SA1 could potentially have an effect on downstream structural elements. I would be particularly interested in the effect of the mutation on SA3 since structure at this region is a well-known splicing regulator and considering the most dramatic changes in the splicing profile occurred to transcripts that produce the *tat* mRNA. This could be an indication that the regulatory structures around the SA3 region are being altered by the changes in structure at SA1. A structural analysis of the SLSA1 mutations in the *vif* transcript would also allow an investigation into any structural change in the nearby splicing enhancer sequences of ESEVif and ESEM1/M2. A change in the structure of these enhancers would affect their recognition by the SR proteins to which they bind.

The structure of the spliced *vpu/env* mRNA also provides possible answers to previously raised questions about the regulation of translation from this transcript including how this single message produces both Env and Vpu proteins. One could speculate that the bicistronic nature of this transcript could be due to the structure of the mRNA, whether it is a ribosomal shunt or an IRES. The mRNA model that I have described, with the Vpu start site located within a single hairpin, suggests a ribosomal shunting pathway wherein the ribosome associates at the 5' cap, disassociates at the stem to bypass the Vpu start codon, then reassociates to recognize the downstream Env AUG site. Many different protocols exist for detection of ribosomal shunting, but the

availability of the RNA secondary structure allows for directed analysis. Mutations to disrupt the stem at the Vpu start codon would allow analysis of whether this structure is acting to disassemble to the ribosome. The SHAPE-determined structure at this region may help to answer the pestering question: What allows Env to be translated from the same message as Vpu?

Lastly, the SHAPE experiments that I performed were limited by the absence of protein in the system, especially concerning the analysis of structure in the HIV-1 mRNA. The regulation of splicing depends on the presence of protein in the nucleus. SR proteins and hnRNPs will bind to specific sequences, and in doing so, strengthen or weaken the surrounding structures. A thorough analysis of these structures as they appear in the cell would include these RNA-binding factors in various concentrations and ratios. The structures that I describe in this work give a snapshot of the RNA prior to protein recognition, but adding binding factors would aide in understanding how the change in these structures may be significant in splicing regulation.

REFERENCES

1. 1997. *In* J. M. Coffin, S. H. Hughes, and H. E. Varmus (ed.), Retroviruses, Cold Spring Harbor (NY).

2. **Abbink, T. E., and B. Berkhout.** 2008. RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor. Journal of virology **82:**3090-3098.

3. **Amendt, B. A., D. Hesslein, L. J. Chang, and C. M. Stoltzfus.** 1994. Presence of negative and positive cis-acting RNA splicing elements within and flanking the first tat coding exon of human immunodeficiency virus type 1. Molecular and cellular biology **14:**3960-3970.

4. **Amendt, B. A., Z. H. Si, and C. M. Stoltzfus.** 1995. Presence of exon splicing silencers within human immunodeficiency virus type 1 tat exon 2 and tat-rev exon 3: evidence for inhibition mediated by cellular factors. Molecular and cellular biology **15:**6480.

5. **Asai, K., C. Platt, and A. Cochrane.** 2003. Control of HIV-1 env RNA splicing and transport: investigating the role of hnRNP A1 in exon splicing silencer (ESS3a) function. Virology **314:**229-242.

6. **Balotta, C., P. Lusso, R. Crowley, R. C. Gallo, and G. Franchini.** 1993. Antisense phosphorothioate oligodeoxynucleotides targeted to the vpr gene inhibit human immunodeficiency virus type 1 replication in primary human macrophages. Journal of virology **67:**4409-4414.

7. **Ban, N., P. Nissen, J. Hansen, M. Capel, P. B. Moore, and T. A. Steitz.** 1999. Placement of protein and RNA structures into a 5 A-resolution map of the 50S ribosomal subunit. Nature **400:**841-847.

8. **Bannwarth, S., and A. Gatignol.** 2005. HIV-1 TAR RNA: the target of molecular interactions between the virus and its host. Curr HIV Res **3:**61-71.

9. **Barash, Y., J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe, and B. J. Frey.** 2010. Deciphering the splicing code. Nature **465:**53-59.

10. **Baudin, F., R. Marquet, C. Isel, J. L. Darlix, B. Ehresmann, and C. Ehresmann.** 1993. Functional sites in the 5' region of human immunodeficiency virus type 1 RNA form defined structural domains. Journal of molecular biology **229:**382-397.

11. **Berget, S. M.** 1995. Exon recognition in vertebrate splicing. The Journal of biological chemistry **270:**2411-2414.

12. **Berkhout, B.** 1992. Structural features in TAR RNA of human and simian immunodeficiency viruses: a phylogenetic analysis. Nucleic acids research **20:**27-31.

13. **Bilodeau, P. S., J. K. Domsic, A. Mayeda, A. R. Krainer, and C. M. Stoltzfus.** 2001. RNA splicing at human immunodeficiency virus type 1 3' splice site A2 is regulated by binding of hnRNP A/B proteins to an exonic splicing silencer element. Journal of virology **75:**8487-8497.

14. **Bishop, K. N., R. K. Holmes, A. M. Sheehy, N. O. Davidson, S. J. Cho, and M. H. Malim.** 2004. Cytidine deamination of retroviral DNA by diverse APOBEC proteins. Current biology : CB **14:**1392-1396.

15. **Bohan, C. A., F. Kashanchi, B. Ensoli, L. Buonaguro, K. A. Boris-Lawrie, and J. N. Brady.** 1992. Analysis of Tat transactivation of human immunodeficiency virus transcription in vitro. Gene expression **2:**391-407.

16. **Bonavia, R., A. Bajetto, S. Barbero, A. Albini, D. M. Noonan, and G. Schettini.** 2001. HIV-1 Tat causes apoptotic death and calcium homeostasis alterations in rat neurons. Biochemical and biophysical research communications **288:**301-308.

17. **Borer, P. N., B. Dengler, I. Tinoco, Jr., and O. C. Uhlenbeck.** 1974. Stability of ribonucleic acid double-stranded helices. Journal of molecular biology **86:**843-853.

18. **Brady, J., and F. Kashanchi.** 2005. Tat gets the "green" light on transcription initiation. Retrovirology **2:**69.

19. **Buratti, E., and F. E. Baralle.** 2004. Influence of RNA secondary structure on the pre-mRNA splicing process. Molecular and cellular biology **24:**10505-10514.

20. **Caputi, M., M. Freund, S. Kammler, C. Asang, and H. Schaal.** 2004. A bidirectional SF2/ASF- and SRp40-dependent splicing enhancer regulates human immunodeficiency virus type 1 rev, env, vpu, and nef gene expression. Journal of virology **78:**6517-6526.

21. **Chang, D. D., and P. A. Sharp.** 1989. Regulation by HIV Rev depends upon recognition of splice sites. Cell **59:**789-795.

22. **Charles Calef, J. M., David H. O'Connor, David I. Watkins, and Bette Korber.** 2001. Numbering Positions in SIV Relative to SIVMM239. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory.

23. **Charneau, P., M. Alizon, and F. Clavel.** 1992. A second origin of DNA plus-strand synthesis is required for optimal human immunodeficiency virus replication. Journal of virology **66:**2814-2820.

24. **Charpentier, B., and M. Rosbash.** 1996. Intramolecular structure in yeast introns aids the early steps of in vitro spliceosome assembly. RNA **2:**509-522.

25. **Chertova, E., J. W. Bess, Jr., B. J. Crise, I. R. Sowder, T. M. Schaden, J. M. Hilburn, J. A. Hoxie, R. E. Benveniste, J. D. Lifson, L. E. Henderson, and L. O. Arthur.** 2002. Envelope glycoprotein incorporation, not shedding of surface envelope glycoprotein (gp120/SU), Is the primary determinant of SU content of purified human immunodeficiency virus type 1 and simian immunodeficiency virus. Journal of virology **76:**5315-5325.

26. **Clever, J. L., and T. G. Parslow.** 1997. Mutant human immunodeficiency virus type 1 genomes with defects in RNA dimerization or encapsidation. Journal of virology **71:**3407-3414.

27. **Cole, J. L., J. D. Gehman, J. A. Shafer, and L. C. Kuo.** 1993. Solution oligomerization of the rev protein of HIV-1: implications for function. Biochemistry **32:**11769-11775.

28. **Connor, R. I., B. K. Chen, S. Choe, and N. R. Landau.** 1995. Vpr is required for efficient replication of human immunodeficiency virus type-1 in mononuclear phagocytes. Virology **206:**935-944.

29. **Conticello, S. G., R. S. Harris, and M. S. Neuberger.** 2003. The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G. Current biology : CB **13:**2009-2013.

30. **Cook, K. S., G. J. Fisk, J. Hauber, N. Usman, T. J. Daly, and J. R. Rusche.** 1991. Characterization of HIV-1 REV protein: binding stoichiometry and minimal RNA substrate. Nucleic acids research **19:**1577-1583.

31. **Costantino, D. A., J. S. Pfingsten, R. P. Rambo, and J. S. Kieft.** 2008. tRNA-mRNA mimicry drives translation initiation from a viral IRES. Nature structural & molecular biology **15:**57-64.

32. **Cullen, B. R.** 2003. Nuclear RNA export. Journal of cell science **116:**587-597.

33. **D'Souza, V., and M. F. Summers.** 2004. Structural basis for packaging the dimeric genome of Moloney murine leukaemia virus. Nature **431:**586-590.

34. **Daly, T. J., K. S. Cook, G. S. Gray, T. E. Maione, and J. R. Rusche.** 1989. Specific binding of HIV-1 recombinant Rev protein to the Rev-responsive element in vitro. Nature **342:**816-819.

35. **Daly, T. J., R. C. Doten, P. Rennert, M. Auer, H. Jaksche, A. Donner, G. Fisk, and J. R. Rusche.** 1993. Biochemical characterization of binding of multiple HIV-1 Rev monomeric proteins to the Rev responsive element. Biochemistry **32:**10497-10505.

36. **Damgaard, C. K., E. S. Andersen, B. Knudsen, J. Gorodkin, and J. Kjems.** 2004. RNA interactions in the 5' region of the HIV-1 genome. Journal of molecular biology **336:**369-379.

37. **Damgaard, C. K., T. O. Tange, and J. Kjems.** 2002. hnRNP A1 controls HIV-1 mRNA splicing through cooperative binding to intron and exon splicing silencers in the context of a conserved secondary structure. RNA **8:**1401-1415.

38. **Daugherty, M. D., I. D'Orso, and A. D. Frankel.** 2008. A solution to limited genomic capacity: using adaptable binding surfaces to assemble the functional HIV Rev oligomer on RNA. Molecular cell **31:**824-834.

39. **Daugherty, M. D., B. Liu, and A. D. Frankel.** 2010. Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. Nature structural & molecular biology **17:**1337-1342.

40. **Deigan, K. E., T. W. Li, D. H. Mathews, and K. M. Weeks.** 2009. Accurate SHAPE-directed RNA structure determination. Proceedings of the National Academy of Sciences of the United States of America **106:**97-102.

41. **Del Gatto-Konczak, F., M. Olive, M. C. Gesnel, and R. Breathnach.** 1999. hnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer. Molecular and cellular biology **19:**251-260.

42. **Domsic, J. K., Y. Wang, A. Mayeda, A. R. Krainer, and C. M. Stoltzfus.** 2003. Human immunodeficiency virus type 1 hnRNP A/B-dependent exonic splicing silencer ESSV antagonizes binding of U2AF65 to viral polypyrimidine tracts. Molecular and cellular biology **23:**8762-8772.

43. **Drake, J. W.** 1993. Rates of spontaneous mutation among RNA viruses. Proceedings of the National Academy of Sciences of the United States of America **90:**4171-4175.

44. **Draper, D. E.** 1996. Strategies for RNA folding. Trends in biochemical sciences **21:**145-149.

45. **Duncan, C. D., and K. M. Weeks.** 2008. SHAPE analysis of long-range interactions reveals extensive and thermodynamically preferred misfolding in a fragile group I intron RNA. Biochemistry **47:**8504-8513.

46. **Dyhr-Mikkelsen, H., and J. Kjems.** 1995. Inefficient spliceosome assembly and abnormal branch site selection in splicing of an HIV-1 transcript in vitro. The Journal of biological chemistry **270:**24060-24066.

47. **Emerman, M., R. Vazeux, and K. Peden.** 1989. The rev gene product of the human immunodeficiency virus affects envelope-specific RNA localization. Cell **57:**1155-1165.

48. **Exline, C. M., Z. Feng, and C. M. Stoltzfus.** 2008. Negative and positive mRNA splicing elements act competitively to regulate human immunodeficiency virus type 1 vif gene expression. Journal of virology **82:**3921-3931.

49. **Felber, B. K., M. Hadzopoulou-Cladaras, C. Cladaras, T. Copeland, and G. N. Pavlakis.** 1989. rev protein of human immunodeficiency virus type 1 affects the stability and transport of the viral mRNA. Proc Natl Acad Sci U S A **86:**1495-1499.

50. **Felden, B., C. Florentz, A. McPherson, and R. Giege.** 1994. A histidine accepting tRNA-like fold at the 3'-end of satellite tobacco mosaic virus RNA. Nucleic acids research **22:**2882-2886.

51. **Fernandez-Miragall, O., S. Lopez de Quinto, and E. Martinez-Salas.** 2009. Relevance of RNA structure for the activity of picornavirus IRES elements. Virus research **139:**172-182.

52. **Fornerod, M., M. Ohno, M. Yoshida, and I. W. Mattaj.** 1997. CRM1 is an export receptor for leucine-rich nuclear export signals. Cell **90:**1051-1060.

53. **Fu, X. D.** 2004. Towards a splicing code. Cell **119:**736-738.

54. **Galli, A., A. Lai, S. Corvasce, F. Saladini, C. Riva, L. Deho, I. Caramma, M. Franzetti, L. Romano, M. Galli, M. Zazzi, and C. Balotta.** 2008. Recombination analysis and structure prediction show correlation between breakpoint clusters and RNA hairpins in the pol gene of human immunodeficiency virus type 1 unique recombinant forms. The Journal of general virology **89:**3119-3125.

55. **Gao, F., E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn.** 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature **397:**436-441.

56. **Gherghe, C., C. W. Leonard, R. J. Gorelick, and K. M. Weeks.** 2010. Secondary structure of the mature ex virio Moloney murine leukemia virus genomic RNA dimerization domain. Journal of virology **84:**898-906.

57. **Grabowski, P. J., F. U. Nasim, H. C. Kuo, and R. Burch.** 1991. Combinatorial splicing of exon pairs by two-site binding of U1 small nuclear ribonucleoprotein particle. Molecular and cellular biology **11:**5919-5928.

58. **Graveley, B. R.** 2000. Sorting out the complexity of SR protein functions. RNA **6:**1197-1211.

59. **Gultyaev, A. P., H. A. Heus, and R. C. Olsthoorn.** 2007. An RNA conformational shift in recent H5N1 influenza A viruses. Bioinformatics **23:**272-276.

60. **Gutell, R. R., B. Weiser, C. R. Woese, and H. F. Noller.** 1985. Comparative anatomy of 16-S-like ribosomal RNA. Progress in nucleic acid research and molecular biology **32:**155-216.

61. **Hallay, H., N. Locker, L. Ayadi, D. Ropers, E. Guittet, and C. Branlant.** 2006. Biochemical and NMR study on the competition between proteins SC35, SRp40, and heterogeneous nuclear ribonucleoprotein A1 at the HIV-1 Tat exon 2 splicing site. The Journal of biological chemistry **281:**37159-37174.

62. **Harrison, G. P., and A. M. Lever.** 1992. The human immunodeficiency virus type 1 packaging signal and major splice donor region have a conserved stable secondary structure. Journal of virology **66:**4144-4153.

63. **Hauber, J., and B. R. Cullen.** 1988. Mutational analysis of the trans-activation-responsive region of the human immunodeficiency virus type I long terminal repeat. Journal of virology **62:**673-679.

64. **Heinzinger, N. K., M. I. Bukinsky, S. A. Haggerty, A. M. Ragland, V. Kewalramani, M. A. Lee, H. E. Gendelman, L. Ratner, M. Stevenson, and M. Emerman.** 1994. The Vpr protein of human immunodeficiency virus type 1 influences nuclear localization of viral nucleic acids in nondividing host cells. Proceedings of the National Academy of Sciences of the United States of America **91:**7311-7315.

65. **Hiller, M., Z. Zhang, R. Backofen, and S. Stamm.** 2007. Pre-mRNA secondary structures influence exon recognition. PLoS genetics **3:**e204.

66. **Holbrook, S. R.** 2005. RNA structure: the long and the short of it. Current opinion in structural biology **15:**302-308.

67. **Holmes, E. C.** 2001. On the origin and evolution of the human immunodeficiency virus (HIV). Biological reviews of the Cambridge Philosophical Society **76:**239-254.

68. **Jacek Nowakowski, I. T.** 1997. RNA Structure and Stability.

69. **Jacks, T., M. D. Power, F. R. Masiarz, P. A. Luciw, P. J. Barr, and H. E. Varmus.** 1988. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. Nature **331:**280-283.

70. **Jacquenet, S., A. Mereau, P. S. Bilodeau, L. Damier, C. M. Stoltzfus, and C. Branlant.** 2001. A second exon splicing silencer within human immunodeficiency virus type 1 tat exon 2 represses splicing of Tat mRNA and binds protein hnRNP H. The Journal of biological chemistry **276:**40464-40475.

71. **Jacquenet, S., D. Ropers, P. S. Bilodeau, L. Damier, A. Mougin, C. M. Stoltzfus, and C. Branlant.** 2001. Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3' splice site and its cis-regulatory element: possible involvement in RNA splicing. Nucleic acids research **29:**464-478.

72. **Jones, C. N., K. A. Wilkinson, K. T. Hung, K. M. Weeks, and L. L. Spremulli.** 2008. Lack of secondary structure characterizes the 5' ends of mammalian mitochondrial mRNAs. RNA **14:**862-871.

73. **Jouvenet, N., S. M. Simon, and P. D. Bieniasz.** 2009. Imaging the interaction of HIV-1 genomes and Gag during assembly of individual viral particles. Proceedings of the National Academy of Sciences of the United States of America **106:**19114-19119.

74. **Jowett, J. B., V. Planelles, B. Poon, N. P. Shah, M. L. Chen, and I. S. Chen.** 1995. The human immunodeficiency virus type 1 vpr gene arrests infected T cells in the G2 + M phase of the cell cycle. Journal of virology **69:**6304-6313.

75. **Kammler, S., C. Leurs, M. Freund, J. Krummheuer, K. Seidel, T. O. Tange, M. K. Lund, J. Kjems, A. Scheid, and H. Schaal.** 2001. The sequence complementarity between HIV-1 5' splice site SD4 and U1 snRNA determines the steady-state level of an unstable env pre-mRNA. RNA **7:**421-434.

76. **Kammler, S., M. Otte, I. Hauber, J. Kjems, J. Hauber, and H. Schaal.** 2006. The strength of the HIV-1 3' splice sites affects Rev function. Retrovirology **3:**89.

77. **Kanki, P. J., K. U. Travers, M. B. S, C. C. Hsieh, R. G. Marlink, N. A. Gueye, T. Siby, I. Thior, M. Hernandez-Avila, J. L. Sankale, and et al.** 1994. Slower heterosexual spread of HIV-2 than HIV-1. Lancet **343:**943-946.

78. **Kao, S., M. A. Khan, E. Miyagi, R. Plishka, A. Buckler-White, and K. Strebel.** 2003. The human immunodeficiency virus type 1 Vif protein reduces intracellular expression and inhibits packaging of APOBEC3G (CEM15), a cellular inhibitor of virus infectivity. Journal of virology **77:**11398-11407.

79. **Karn, J.** 1999. Tackling Tat. Journal of molecular biology **293:**235-254.

80. **Karn, J., and C. M. Stoltzfus.** 2012. Transcriptional and Posttranscriptional Regulation of HIV-1 Gene Expression. Cold Spring Harbor perspectives in medicine **2:**a006916.

81. **Kestler, H., T. Kodama, D. Ringler, M. Marthas, N. Pedersen, A. Lackner, D. Regier, P. Sehgal, M. Daniel, N. King, and et al.** 1990. Induction of AIDS in rhesus monkeys by molecularly cloned simian immunodeficiency virus. Science **248:**1109-1112.

82. **Kimpton, J., and M. Emerman.** 1992. Detection of replication-competent and pseudotyped human immunodeficiency virus with a sensitive cell line on the basis of activation of an integrated beta-galactosidase gene. Journal of virology **66:**2232-2239.

83. **Kjems, J., M. Brown, D. D. Chang, and P. A. Sharp.** 1991. Structural analysis of the interaction between the human immunodeficiency virus Rev protein and the Rev response element. Proceedings of the National Academy of Sciences of the United States of America **88:**683-687.

84. **Kjems, J., A. D. Frankel, and P. A. Sharp.** 1991. Specific regulation of mRNA splicing in vitro by a peptide from HIV-1 Rev. Cell **67:**169-178.

85. **Kleiman, L.** 2002. tRNA(Lys3): the primer tRNA for reverse transcription in HIV-1. IUBMB life **53:**107-114.

86. **Knies, J. L., K. K. Dang, T. J. Vision, N. G. Hoffman, R. Swanstrom, and C. L. Burch.** 2008. Compensatory evolution in RNA secondary structures increases substitution rate variation among sites. Molecular biology and evolution **25:**1778-1787.

87. **Koenig, R., S. Barends, A. P. Gultyaev, D. E. Lesemann, H. J. Vetten, S. Loss, and C. W. Pleij.** 2005. Nemesia ring necrosis virus: a new tymovirus with a genomic RNA having a histidylatable tobamovirus-like 3' end. The Journal of general virology **86:**1827-1833.

88. **Kollmus, H., A. Honigman, A. Panet, and H. Hauser.** 1994. The sequences of and distance between two cis-acting signals determine the efficiency of ribosomal frameshifting in human immunodeficiency virus type 1 and human T-cell leukemia virus type II in vivo. Journal of virology **68:**6087-6091.

89. **Kramer, W., V. Drutsa, H. W. Jansen, B. Kramer, M. Pflugfelder, and H. J. Fritz.** 1984. The gapped duplex DNA approach to oligonucleotide-directed mutation construction. Nucleic acids research **12:**9441-9456.

90. **Kruger, K., P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech.** 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell **31:**147-157.

91. **Kutluay, S. B., and P. D. Bieniasz.** 2010. Analysis of the initiating events in HIV-1 particle assembly and genome packaging. PLoS pathogens **6:**e1001200.

92. **Lanciault, C., and J. J. Champoux.** 2006. Pausing during reverse transcription increases the rate of retroviral recombination. Journal of virology **80:**2483-2494.

93. **Laspia, M. F., A. P. Rice, and M. B. Mathews.** 1989. HIV-1 Tat protein increases transcriptional initiation and stabilizes elongation. Cell **59:**283-292.

94. **Le, S. Y., J. H. Chen, D. Chatterjee, and J. V. Maizel.** 1989. Sequence divergence and open regions of RNA secondary structures in the envelope regions of the 17 human immunodeficiency virus isolates. Nucleic acids research **17:**3275-3288.

95. **Legrain, P., and M. Rosbash.** 1989. Some cis- and trans-acting mutants for splicing target pre-mRNA to the cytoplasm. Cell **57:**573-583.

96. **Levengood, J. D., C. Rollins, C. H. Mishler, C. A. Johnson, G. Miner, P. Rajan, B. M. Znosko, and B. S. Tolbert.** 2012. Solution structure of the HIV-1 exon splicing silencer 3. Journal of molecular biology **415:**680-698.

97. **Li, C. J., D. J. Friedman, C. Wang, V. Metelev, and A. B. Pardee.** 1995. Induction of apoptosis in uninfected lymphocytes by HIV-1 Tat protein. Science **268:**429-431.

98. **Liu, W., Y. Zhou, Z. Hu, T. Sun, A. Denise, X. D. Fu, and Y. Zhang.** 2010. Regulation of splicing enhancer activities by RNA secondary structures. FEBS letters **584:**4401-4407.

99. **Low, J. T., S. A. Knoepfel, J. M. Watts, O. ter Brake, B. Berkhout, and K. M. Weeks.** 2012. SHAPE-directed discovery of potent shRNA inhibitors of HIV-1. Molecular therapy : the journal of the American Society of Gene Therapy **20:**820-828.

100. **Lu, K., X. Heng, L. Garyu, S. Monti, E. L. Garcia, S. Kharytonchyk, B. Dorjsuren, G. Kulandaivel, S. Jones, A. Hiremath, S. S. Divakaruni, C. LaCotti, S. Barton, D. Tummillo, A. Hosic, K. Edme, S. Albrecht, A. Telesnitsky, and M. F. Summers.** 2011. NMR detection of structures in the HIV-1 5'-leader RNA that regulate genome packaging. Science **334:**242-245.

101. **Lu, K., X. Heng, and M. F. Summers.** 2011. Structural determinants and mechanism of HIV-1 genome packaging. Journal of molecular biology **410:**609-633.

102. **Madhani, H. D., and C. Guthrie.** 1994. Dynamic RNA-RNA interactions in the spliceosome. Annual review of genetics **28:**1-26.

103. **Madsen, J. M., and C. M. Stoltzfus.** 2005. An exonic splicing silencer downstream of the 3' splice site A2 is required for efficient human immunodeficiency virus type 1 replication. Journal of virology **79:**10478-10486.

104. **Madsen, J. M., and C. M. Stoltzfus.** 2006. A suboptimal 5' splice site downstream of HIV-1 splice site A1 is required for unspliced viral mRNA accumulation and efficient virus replication. Retrovirology **3:**10.

105. **Malim, M. H., S. Bohnlein, J. Hauber, and B. R. Cullen.** 1989. Functional dissection of the HIV-1 Rev trans-activator--derivation of a trans-dominant repressor of Rev function. Cell **58:**205-214.

106. **Malim, M. H., J. Hauber, S. Y. Le, J. V. Maizel, and B. R. Cullen.** 1989. The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. Nature **338:**254-257.

107. **Mandal, D., C. M. Exline, Z. Feng, and C. M. Stoltzfus.** 2009. Regulation of Vif mRNA splicing by human immunodeficiency virus type 1 requires 5' splice site D2 and an exonic splicing enhancer to counteract cellular restriction factor APOBEC3G. Journal of virology **83:**6067-6078.

108. **Mandal, D., Z. Feng, and C. M. Stoltzfus.** 2008. Gag-processing defect of human immunodeficiency virus type 1 integrase E246 and G247 mutants is caused by activation of an overlapping 5' splice site. Journal of virology **82:**1600-1604.

109. **Mandal, M., and R. R. Breaker.** 2004. Gene regulation by riboswitches. Nature reviews. Molecular cell biology **5:**451-463.

110. **Mann, D. A., I. Mikaelian, R. W. Zemmel, S. M. Green, A. D. Lowe, T. Kimura, M. Singh, P. J. Butler, M. J. Gait, and J. Karn.** 1994. A molecular rheostat. Co-operative rev binding to stem I of the rev-response element modulates human immunodeficiency virus type-1 late gene expression. Journal of molecular biology **241:**193-207.

111. **Marchand, V., A. Mereau, S. Jacquenet, D. Thomas, A. Mougin, R. Gattoni, J. Stevenin, and C. Branlant.** 2002. A Janus splicing regulatory element modulates HIV-1 tat and rev mRNA production by coordination of hnRNP A1 cooperative binding. Journal of molecular biology **323:**629-652.

112. **Marin, M., K. M. Rose, S. L. Kozak, and D. Kabat.** 2003. HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. Nature medicine **9:**1398-1403.

113. **Marlink, R., P. Kanki, I. Thior, K. Travers, G. Eisen, T. Siby, I. Traore, C. C. Hsieh, M. C. Dia, E. H. Gueye, and et al.** 1994. Reduced rate of disease development after HIV-2 infection as compared to HIV-1. Science **265:**1587-1590.

114. **Mathews, D. H., M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner.** 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proceedings of the National Academy of Sciences of the United States of America **101:**7287-7292.

115. **Matlin, A. J., F. Clark, and C. W. Smith.** 2005. Understanding alternative splicing: towards a cellular code. Nature reviews. Molecular cell biology **6:**386-398.

116. **Matta, C. F., N. Castillo, and R. J. Boyd.** 2006. Extended weak bonding interactions in DNA: pi-stacking (base-base), base-backbone, and backbone-backbone interactions. The journal of physical chemistry. B **110:**563-578.

117. **Means, R. E., T. Matthews, J. A. Hoxie, M. H. Malim, T. Kodama, and R. C. Desrosiers.** 2001. Ability of the V3 loop of simian immunodeficiency virus to serve as a target for antibody-mediated neutralization: correlation of neutralization sensitivity, growth in macrophages, and decreased dependence on CD4. Journal of virology **75:**3903-3915.

118. **Merino, E. J., K. A. Wilkinson, J. L. Coughlan, and K. M. Weeks.** 2005. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). Journal of the American Chemical Society **127:**4223-4231.

119. **Mignon, P., S. Loverix, J. Steyaert, and P. Geerlings.** 2005. Influence of the pi-pi interaction on the hydrogen bonding capacity of stacked DNA/RNA bases. Nucleic acids research **33:**1779-1789.

120. **Mir, M. A., B. Brown, B. Hjelle, W. A. Duran, and A. T. Panganiban.** 2006. Hantavirus N protein exhibits genus-specific recognition of the viral RNA panhandle. Journal of virology **80:**11283-11292.

121. **Moore, M. D., O. A. Nikolaitchik, J. Chen, M. L. Hammarskjold, D. Rekosh, and W. S. Hu.** 2009. Probing the HIV-1 genomic RNA trafficking pathway and dimerization by genetic recombination and single virion analyses. PLoS pathogens **5:**e1000627.

122. **Mortimer, S. A., and K. M. Weeks.** 2007. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. Journal of the American Chemical Society **129:**4144-4145.

123. **Muesing, M. A., D. H. Smith, C. D. Cabradilla, C. V. Benton, L. A. Lasky, and D. J. Capon.** 1985. Nucleic acid structure and expression of the human AIDS/lymphadenopathy retrovirus. Nature **313:**450-458.

124. **Muesing, M. A., D. H. Smith, and D. J. Capon.** 1987. Regulation of mRNA accumulation by a human immunodeficiency virus trans-activator protein. Cell **48:**691-701.

125. **Nasim, F. U., S. Hutchison, M. Cordeau, and B. Chabot.** 2002. High-affinity hnRNP A1 binding sites and duplex-forming inverted repeats have similar effects on 5' splice site selection in support of a common looping out and repression mechanism. RNA **8:**1078-1089.

126. **Nissen, P., J. A. Ippolito, N. Ban, P. B. Moore, and T. A. Steitz.** 2001. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. Proceedings of the National Academy of Sciences of the United States of America **98:**4899-4903.

127. **Noller, H. F.** 2005. RNA structure: reading the ribosome. Science **309:**1508-1514.

128. **Noller, H. F.** 1984. Structure of ribosomal RNA. Annual review of biochemistry **53:**119-162.

129. **O'Reilly, M. M., M. T. McNally, and K. L. Beemon.** 1995. Two strong 5' splice sites and competing, suboptimal 3' splice sites involved in alternative splicing of human immunodeficiency virus type 1 RNA. Virology **213:**373-385.

130. **Olsen, H. S., P. Nelbock, A. W. Cochrane, and C. A. Rosen.** 1990. Secondary structure is the major determinant for interaction of HIV rev protein with RNA. Science **247:**845-848.

131. **Paillart, J. C., M. Dettenhofer, X. F. Yu, C. Ehresmann, B. Ehresmann, and R. Marquet.** 2004. First snapshots of the HIV-1 RNA structure in infected cells and in virions. The Journal of biological chemistry **279:**48397-48403.

132. **Paillart, J. C., R. Marquet, E. Skripkin, B. Ehresmann, and C. Ehresmann.** 1994. Mutational analysis of the bipartite dimer linkage structure of human immunodeficiency virus type 1 genomic RNA. The Journal of biological chemistry **269:**27486-27493.

133. **Paillart, J. C., M. Shehu-Xhilaga, R. Marquet, and J. Mak.** 2004. Dimerization of retroviral RNA genomes: an inseparable pair. Nature reviews. Microbiology **2:**461-472.

134. **Paillart, J. C., E. Skripkin, B. Ehresmann, C. Ehresmann, and R. Marquet.** 2002. In vitro evidence for a long range pseudoknot in the 5'-untranslated and matrix coding regions of HIV-1 genomic RNA. The Journal of biological chemistry **277:**5995-6004.

135. **Pan, Q., O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe.** 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature genetics **40:**1413-1415.

136. **Parkin, N. T., M. Chamorro, and H. E. Varmus.** 1992. Human immunodeficiency virus type 1 gag-pol frameshifting is dependent on downstream mRNA secondary structure: demonstration by expression in vivo. Journal of virology **66:**5147-5151.

137. **Pedersen, J. S., I. M. Meyer, R. Forsberg, P. Simmonds, and J. Hein.** 2004. A comparative method for finding and folding RNA secondary structures within protein-coding regions. Nucleic acids research **32:**4925-4936.

138. **Pfingsten, J. S., and J. S. Kieft.** 2008. RNA structure-based ribosome recruitment: lessons from the Dicistroviridae intergenic region IRESes. RNA **14:**1255-1263.

139. **Pollard, V. W., and M. H. Malim.** 1998. The HIV-1 Rev protein. Annual review of microbiology **52:**491-532.

140. **Poole, E., P. Strappe, H. P. Mok, R. Hicks, and A. M. Lever.** 2005. HIV-1 Gag-RNA interaction occurs at a perinuclear/centrosomal site; analysis by confocal microscopy and FRET. Traffic **6:**741-755.

141. **Post, K., B. Kankia, S. Gopalakrishnan, V. Yang, E. Cramer, P. Saladores, R. J. Gorelick, J. Guo, K. Musier-Forsyth, and J. G. Levin.** 2009. Fidelity of plus-strand priming requires the nucleic acid chaperone activity of HIV-1 nucleocapsid protein. Nucleic acids research **37:**1755-1766.

142. **Purcell, D. F., and M. A. Martin.** 1993. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. Journal of virology **67:**6365-6378.

143. **Reed, R.** 1996. Initial splice-site recognition and pairing during pre-mRNA splicing. Current opinion in genetics & development **6:**215-220.

144. **Regier, D. A., and R. C. Desrosiers.** 1990. The complete nucleotide sequence of a pathogenic molecular clone of simian immunodeficiency virus. AIDS research and human retroviruses **6:**1221-1231.

145. **Resch, W., N. Parkin, E. L. Stuelke, T. Watkins, and R. Swanstrom.** 2001. A multiple-site-specific heteroduplex tracking assay as a tool for the study of viral population dynamics. Proceedings of the National Academy of Sciences of the United States of America **98:**176-181.

146. **Reuter, J. S., and D. H. Mathews.** 2010. RNAstructure: software for RNA secondary structure prediction and analysis. BMC bioinformatics **11:**129.

147. **Robberson, B. L., G. J. Cote, and S. M. Berget.** 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. Molecular and cellular biology **10:**84-94.

148. **Saenger, W.** 1984. Principles of Nucleic Acid Structure. Springer-Verlag, New York.

149. **Saliou, J. M., C. F. Bourgeois, L. Ayadi-Ben Mena, D. Ropers, S. Jacquenet, V. Marchand, J. Stevenin, and C. Branlant.** 2009. Role of RNA structure and protein factors in the control of HIV-1 splicing. Frontiers in bioscience : a journal and virtual library **14:**2714-2729.

150. **Schroeder, R., R. Grossberger, A. Pichler, and C. Waldsich.** 2002. RNA folding in vivo. Current opinion in structural biology **12:**296-300.

151. **Sheehy, A. M., N. C. Gaddis, and M. H. Malim.** 2003. The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. Nature medicine **9:**1404-1407.

152. **Shepard, P. J., and K. J. Hertel.** 2008. Conserved RNA secondary structures promote alternative splicing. RNA **14:**1463-1469.

153. **Si, Z., B. A. Amendt, and C. M. Stoltzfus.** 1997. Splicing efficiency of human immunodeficiency virus type 1 tat RNA is determined by both a suboptimal 3' splice site and a 10 nucleotide exon splicing silencer element located within tat exon 2. Nucleic acids research **25:**861-867.

154. **Sinck, L., D. Richer, J. Howard, M. Alexander, D. F. Purcell, R. Marquet, and J. C. Paillart.** 2007. In vitro dimerization of human immunodeficiency virus type 1 (HIV-1) spliced RNAs. RNA **13:**2141-2150.

155. **Skripkin, E., J. C. Paillart, R. Marquet, B. Ehresmann, and C. Ehresmann.** 1994. Identification of the primary site of the human immunodeficiency virus type 1 RNA dimerization in vitro. Proceedings of the National Academy of Sciences of the United States of America **91:**4945-4949.

156. **Smith, C. W., and J. Valcarcel.** 2000. Alternative pre-mRNA splicing: the logic of combinatorial control. Trends in biochemical sciences **25:**381-388.

157. **Staffa, A., and A. Cochrane.** 1994. The tat/rev intron of human immunodeficiency virus type 1 is inefficiently spliced because of suboptimal signals in the 3' splice site. Journal of virology **68:**3071-3079.

158. **Stein, B. S., and E. G. Engleman.** 1990. Intracellular processing of the gp160 HIV-1 envelope precursor. Endoproteolytic cleavage occurs in a cis or medial compartment of the Golgi complex. The Journal of biological chemistry **265:**2640-2649.

159. **Steitz, T. A., and P. B. Moore.** 2003. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Trends in biochemical sciences **28:**411-418.

160. **Stoltzfus, C. M.** 2009. Chapter 1. Regulation of HIV-1 alternative RNA splicing and its role in virus replication. Advances in virus research **74:**1-40.

161. **Swanstrom, R., W. J. DeLorbe, J. M. Bishop, and H. E. Varmus.** 1981. Nucleotide sequence of cloned unintegrated avian sarcoma virus DNA: viral DNA contains direct and inverted repeats similar to those in transposable elements. Proceedings of the National Academy of Sciences of the United States of America **78:**124-128.

162. **Tan, R., L. Chen, J. A. Buettner, D. Hudson, and A. D. Frankel.** 1993. RNA recognition by an isolated alpha helix. Cell **73:**1031-1040.

163. **Tejedor, J. R., and J. Valcarcel.** 2010. Gene regulation: Breaking the second genetic code. Nature **465:**45-46.

164. **Vasa, S. M., N. Guex, K. A. Wilkinson, K. M. Weeks, and M. C. Giddings.** 2008. ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. RNA **14:**1979-1990.

165. **Victoria, J. G., and W. E. Robinson, Jr.** 2005. Disruption of the putative splice acceptor site for SIV(mac239)Vif reveals tight control of SIV splicing and impaired replication in Vif non-permissive cells. Virology **338:**281-291.

166. **Wahl, M. C., C. L. Will, and R. Luhrmann.** 2009. The spliceosome: design principles of a dynamic RNP machine. Cell **136:**701-718.

167. **Wang, Q., I. Barr, F. Guo, and C. Lee.** 2008. Evidence of a novel RNA secondary structure in the coding region of HIV-1 pol gene. RNA **14:**2478-2488.

168. **Wang, S., L. Mortazavi, and K. A. White.** 2008. Higher-order RNA structural requirements and small-molecule induction of tombusvirus subgenomic mRNA transcription. Journal of virology **82:**3864-3871.

169. **Warf, M. B., and J. A. Berglund.** 2010. Role of RNA structure in regulating pre-mRNA splicing. Trends in biochemical sciences **35:**169-178.

170. **Watanabe, T., and B. A. Sullenger.** 2000. Induction of wild-type p53 activity in human cancer cells by ribozymes that repair mutant p53 transcripts. Proceedings of the National Academy of Sciences of the United States of America **97:**8490-8494.

171. **Watkins, T., W. Resch, D. Irlbeck, and R. Swanstrom.** 2003. Selection of high-level resistance to human immunodeficiency virus type 1 protease inhibitors. Antimicrobial agents and chemotherapy **47:**759-769.

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

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

174. **Wlodawer, A., and A. Gustchina.** 2000. Structural and biochemical studies of retroviral proteases. Biochimica et biophysica acta **1477:**16-34.

175. **Yu, X., Y. Yu, B. Liu, K. Luo, W. Kong, P. Mao, and X. F. Yu.** 2003. Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. Science **302:**1056-1060.

176. **Zahler, A. M., C. K. Damgaard, J. Kjems, and M. Caputi.** 2004. SC35 and heterogeneous nuclear ribonucleoprotein A/B proteins bind to a juxtaposed exonic splicing enhancer/exonic splicing silencer element to regulate HIV-1 tat exon 2 splicing. The Journal of biological chemistry **279:**10077-10084.

177. **Zhang, G., M. L. Zapp, G. Yan, and M. R. Green.** 1996. Localization of HIV-1 RNA in mammalian nuclei. The Journal of cell biology **135:**9-18.

178. **Zhang, J., C. C. Kuo, and L. Chen.** 2011. GC content around splice sites affects splicing through pre-mRNA secondary structures. BMC genomics **12:**90.

179. **Zuker, M.** 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research **31:**3406-3415.