Next-Generation Sequencing Techniques to Study HIV-1 Transcription and RNA Structure

NEXT-GENERATION SEQUENCING TECHNIQUES TO STUDY HIV-1 TRANSCRIPTION AND RNA STRUCTURE

A dissertation presented

by

Phillip Tomezsko

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

In the subject of

Virology

Harvard University

Cambridge, Massachusetts

May 2020

Dissertation Advisors: Daniel Kuritzkes & Athe Tsibris Phillip Tomezsko

Next-generation sequencing techniques to study HIV-1 transcription and RNA structure

Abstract

HIV-1 remains a global health challenge. New sequencing techniques and bioinformatic approaches are being developed to study key aspects of HIV-1 biology. I leveraged and developed innovative deep-sequencing approaches to advance our understanding of HIV-1 RNA structure and latency.

In Chapter 1, I will review the HIV-1 replication cycle. Special focus will be devoted to the roles that HIV-1 RNA structures play during replication. I will also review HIV-1 splicing and the current understating of its regulation. Understanding the splice pattern of HIV-1 is crucial to both projects, as it affects the interpretation of RNAseq data. I will then review the latest research on HIV-1 latency, which has been the focus of intense research.

Measuring RNA structure by chemical probing and deep sequencing is a complex technique that must be developed and adapted for each experimental system. In Chapter 2, the technical development of Dimethyl Sulfate (DMS)- Mutational Profiling and Sequencing (MaPseq) for HIV-1 infected cells and HIV-1 virions will be described.

Few studies analyze intracellular RNA structure and no previous technique can measure alternate RNA structure in cells. To address these knowledge gaps, I utilized a novel alternate RNA structure detecting algorithm in conjunction with DMS-MaPseq. In Chapter 3, the use of this

system to study the Rev response element (RRE) within cells, to identify novel RNA structures that regulate splicing, and to measure overall HIV-1 RNA structure heterogeneity will be presented and discussed.

The mechanisms that regulate HIV-1 transcriptional latency and reactivation remain incompletely understood. In Chapter 4, I will discuss our study that developed an enrichment-based RNAseq technique to study HIV-1 reversal of latency in response to a number of drug candidates. We were able to detect, measure, and quantify coverage across the HIV-1 genome despite the extreme rarity of HIV-1 RNA.

Finally, in Chapter 5, I will discuss how these techniques can be further used to advance the study of basic HIV-1 biology and how the techniques can be used to help develop better latency reversing agents. I will also speculate on the role that RNA structure may contribute to HIV-1 latency.

Acknowledgments

I would to thank Daniel Kuritzkes for all of his mentorship during my PhD. I have been unceasingly impressed by his dedication, breadth of knowledge and ability to process data. I have really appreciated the flexible and collaborative environment he has set up.

Athe Tsibris has taught me a tremendous amount about how to think big and come up with bold proposals. I am extremely thankful for his mentorship and leadership during my graduate school career. He gave me a lot of opportunities to grow and put tremendous trust in the data I produced.

I want to thank Silvi Rouskin for taking a chance on collaborating with me, and ending up becoming a co-mentor. Silvi is one of the most creative and dedicated scientists I have met. I am excited for her to start the next step in her journey and I am proud that I was able to help her along the way.

Thanks to the members of my dissertation advisory committee, Alan Engelman, Victoria D'Souza and Benjamin Gewurz. Their guidance over the years was indispensable and I valued it highly.

Two brave virology students led the way for me in the Kuritzkes lab, Fernanda Ferriera and Radwa Sharaf. Thanks to both of them for all the laughs, insightful conversations and support they gave me throughout my PhD.

I want to thank the entire Virology program- students, faculty and administrators. I couldn't imagine a more interesting group of people with whom to learn and work. And a special shout-out to the post-data club magic crew.

I would like to thank Paromita, Jay, Harish, Sitara, Margalit, and Tammy for adopting me into their lab and allowing me to crash the Whitehead retreat, I'll never forget Legends 1291.

A thanks to Kendyll and Heather for all the laughs, some of the tears (that google commercial) and all the fridge wine.

To Stephanie, Olivia and Qianjing, thanks for being terrific lab mates and friends. Pass the P200!

Our lab is held together by Francoise, to whom I am grateful for her tireless effort as the lab manager. Her dedication is inspiring and she never failed to make me laugh.

Thanks to the members of the Li and Lichterfeld labs for their help in lab meetings and advice on scientific matters. And for the fantasy football leagues, paint nights and general lab shenanigans. A special thanks to Zixin Hu for being the first person in the lab to show me the ropes.

I want to thank my Mom and Dad for all the support they have given me. I appreciate how hard they worked to give me a great education and was strengthened by the confidence they have had in me.

I also want to thank my siblings, Greg, Diana, Mat and Tom, for their love and support.

To my nieces and nephew, Cate, Rome Annie and Jovan, who were born during my graduate school career, thank you for inspiring me to work for a better future.

Anna, I can't thank you enough for sticking with me through the thick and thin. Will you marry me?

Lastly, I would like to thank all the blood donors and study participants that made these projects possible.

Table of Contents

Chapter 1: Introduction

1.1 Introduction

Human immunodeficiency virus (HIV)-1 is a retrovirus that infects 37.9 million people worldwide as of 2018[1]. HIV-1 is the causative agent of acquired immunodeficiency syndrome (AIDS). There have been many successes in the treatment and prevention of HIV-1. New infections are down 40% from the peak in 1997 and AIDS-related deaths are down 56% from the peak in 2004[1]. Furthermore, over 23 million people are currently on antiretroviral therapy (ART). However, there is still significant progress to be made. Only 53% of the people living with HIV-1 are fully suppressed on ART and there are over 1 million new infections per year[1]. We still do not have a vaccine or cure for HIV-1. In order to develop the next generation of tools to treat and prevent HIV-1 infection, gaining new insight into HIV-1 using the most cutting-edge technology is critical.

The revolution in next-generation sequencing provides an immense opportunity advance the understanding of HIV-1 biology. Sequencing technology has rapidly changed the applications of nucleic acid sequencing in biological research. New sequencing techniques and bioinformatic approaches are being developed to study key aspects of HIV-1 biology. In my dissertation work, I have focused on two main projects. First, I have adapted Dimethyl Sulfate Mutational Profiling and Sequencing (DMS-MaPseq) to define HIV-1 RNA structure rigorously and applied a novel clustering algorithm to identify RNA structures that regulate HIV-1 splicing. Second, to assess HIV-1 transcription after reactivation from latency, I have developed a novel RNAseq-based enrichment approach that accurately and reproducibly quantifies virus transcripts in primary resting $CD4^+$ T cells from antiretroviral suppressed participants with HIV.

1.2 HIV-1 Replication Cycle

Entry

The first step in the HIV-1 replication cycle is entry into the target cell (see Figure 1.1). HIV-1 predominately infects $CD4^+$ T cells. The envelope (Env) protein forms a trimer that undergoes a conformational shift upon binding of CD4 and a co-receptor on the target cell surface. The conformation change exposes the fusion peptide and facilities membrane-membrane fusion. The main co-receptor is CCR5 but some strains use CXCR4. Macrophages and other myeloid cells can be infected by HIV-1, although at a lower frequency than CD4+ T cells. The core particle, which contains two copies of the ~10 kb RNA genome along with the virally encoded enzymes reverse transcriptase (RT) and integrase (IN) and certain virally encoded accessory proteins such as Vpr, is deposited into the cell's cytoplasm after membrane fusion[2].

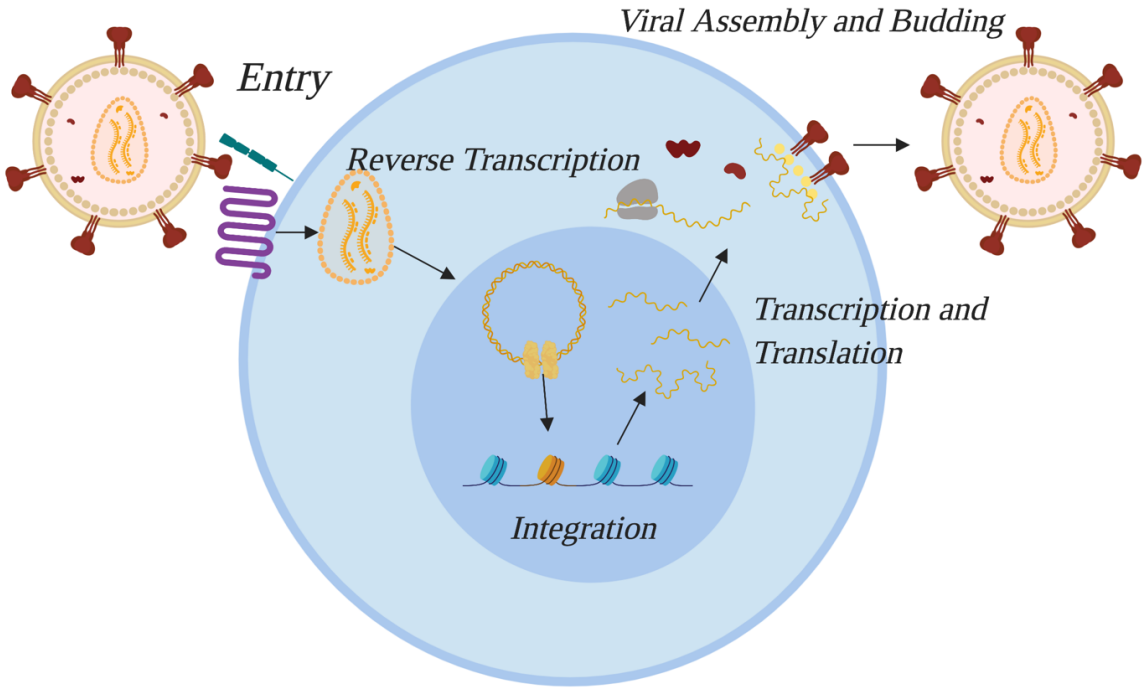

***Figure 1.1**: HIV-1 replication cycle.*

Reverse Transcription

Reverse transcription is a multi-step process that results in the production of a double-stranded DNA HIV-1 genome from the RNA genome that can be integrated into the host genome. The process is carried out by the error-prone RT encoded by the virus. Reverse transcription begins at the primer binding site (PBS) of the 5'UTR, using the $tRNA^{Lys3}$ as a primer[3]. The short DNA reverse transcription product then transfers strands and primes the 3'UTR from the complementary repeat element R. The (-) strand is reverse transcribed to the 5'LTR PBS. The RNA template is degraded by the RNase H activity of RT except for two small portions of degradation-resistant sequence called the polypurine tracts on the (+) strand. These small sequences of RNA remain bound and allow for reverse transcription of the (+) strand. A second strand transfer occurs based on the complementarity of the PBS. DNA synthesis occurs 5' to 3' for both strands simultaneously, in order to synthesize the rest of the 5'LTR for the (-) strand and the entirety of the (+) strand. This whole process is reviewed and more thoroughly described in references[4-6]. Reverse transcription is the source of much of the genetic diversity for HIV-1 through two mechanisms. The first mechanism is that the RT enzyme is error-prone, and makes an average of one mutation per ~7000 nt in cells[7,8]. The second mechanism is recombination of the two packaged genomes during the first strand transfer. This second mechanism is dependent on packaging of non-identical genomes in the same virion, which can happen if a cell contains multiple intact proviruses. The RT is also able to perform inter-strand jumps during either the first or second strand transfer, resulting in deletions or inversions. Additional significant contributions to HIV-1 genetic diversity come from the host RNA polymerase II and a family of cytidine deaminases[9].

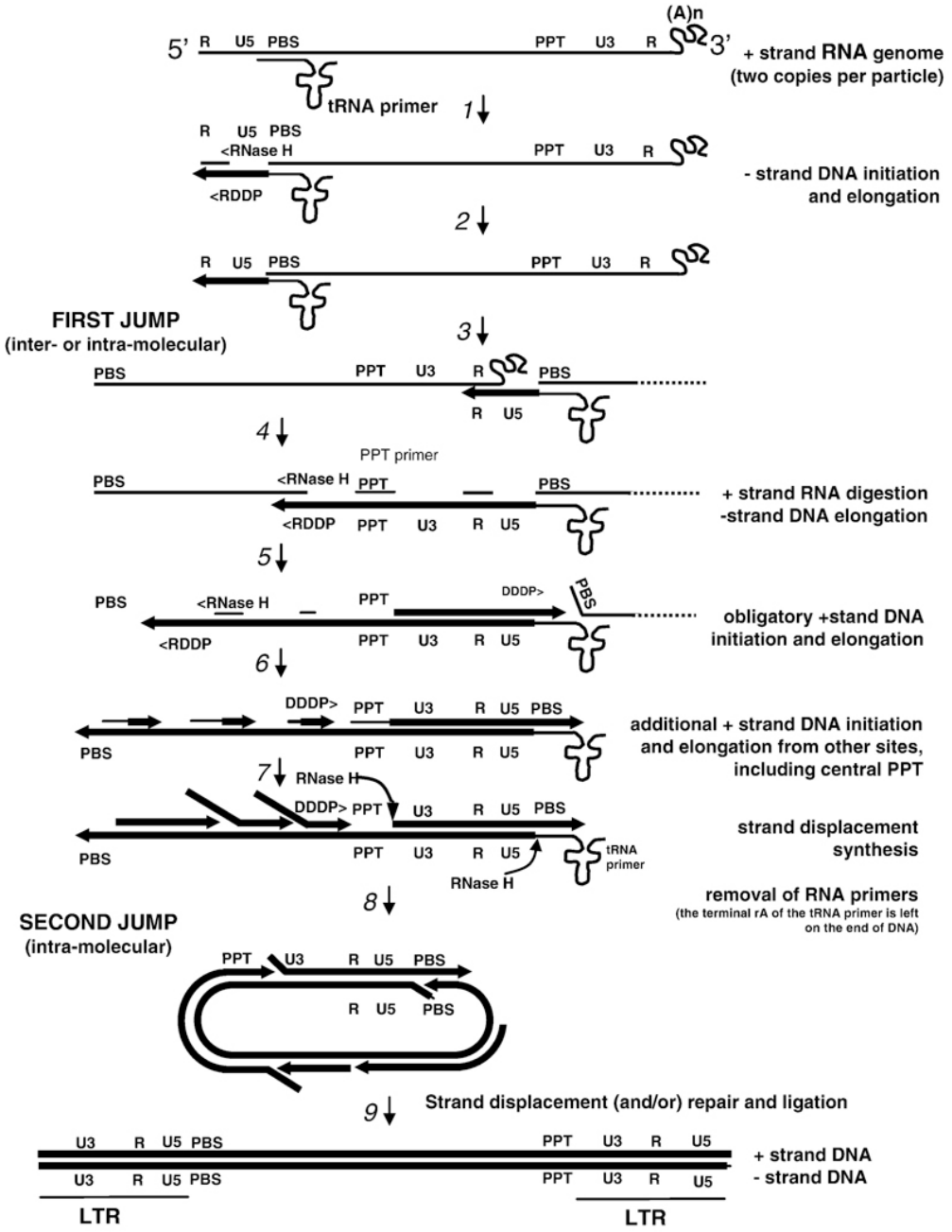

Figure 1.2*: HIV-1 reverse transcription process. Figure reproduced from Sarafianos et al.*[10].

Integration

The linear double-stranded DNA reverse transcription product binds a multimer of the viral enzyme IN to form the pre-integration complex (PIC). The PIC is transported to the nucleus, mediated by an interaction between the HIV-1 capsid protein (CA) and the human protein

CPSF6[11,12]. CPSF6 allows the PIC to interact with a number of human nuclear import factors, including TNPO3[13]. The combined action of these host factors allows for the nuclear import of the PIC near transcriptionally active units of chromatin with a high density of genes[14]. Downstream of nuclear import, the PIC interacts with the tethering factor LEDGF/p75, which is responsible for integration site-selection, most frequently into host genes undergoing active transcription[14-19]. IN has two enzymatic activities; 3' processing and strand transfer. Durin 3' processing, IN hydrolyzes the DNA ends adjacent to the invariant CA sequence. This processing generally removes a dinucleotide from each 3' end. Next, the IN combines the viral and human DNA during strand transfer by using the CA-OH ends to cut host DNA with a 4-6 bp stagger. This action concomitantly joins the viral DNA ends to host DNA 5' phosphates[20].

HIV-1 Transcription and translation

During production infection, the most important driver of HIV-1 transcription is the viral protein Tat. Tat interacts with the viral RNA trans-activation response (TAR) element in order to recruit the human RNA polymerase elongation factor PTEF-b to the viral promoter[21-25]. As more viral RNA is produced, more Tat is able to recruit the elongation factor, forming a positive feedback loop. Efficient promotion of HIV-1 RNA is also dependent on the nuclear localization of several human transcription factors for initiation, Sp1, AP-1, NF-κB and NFAT[26-29]. Several of these transcription factors become localized to the nucleus during T cell activation. After transcription, the HIV-1 RNA is either singly spliced (SS), multiply spliced (MS) or unspliced (US). These species of viral RNAs are referred to as size classes[30,31].

The accessory proteins genes Vif, Vpr and Tat are produced from both SS and MS transcripts. Rev and Nef are translated from MS transcripts exclusively. The start codon for each gene is set by the splice acceptor site that is used. Vpu/Env are both translated from the same transcript, which is exclusively MS. Vpu is translated from an upstream open reading frame (uORF) of Env. Gag and Pol are translated from the unspliced RNA[31]. Pol is only translated as a result of a ribosome -1 frameshift that allows the normal Gag stop codon to be read through[32,33].

Viral assembly and budding

Assembly of the virion occurs at the plasma membrane of the cell. The most important viral protein for this process is the Gag polyprotein, which consists of the matrix protein, capsid protein, nucleocapsid protein and scaffold protein p6. N-terminal myristylation of Gag directs the Gag polyprotein to the cell membrane, which then results in concentration of Env at the cell membrane and recruits genomic RNA into the budding virion. Gag polyprotein can form the spherical core of the immature virion however it requires the host endosomal sorting complexes required for transport (ESCRT) machinery in order to catalyze the membrane fission needed for the virion to bud from the cell. After budding, HIV-1 protease is activated and cleaves the Gag-PR-RT-IN and the Gag polyproteins into their constitutive elements to form the mature virion[34].

1.3 RNA structure and the viral replication cycle

One way for RNA viruses to maximize the coding capacity of a short genome is to employ RNA structures to regulate certain steps in the viral replication cycle. The study of RNA structure has been exceptionally difficult due to the multitude of base-pairing combinations that exist for each RNA. The base-pairing combinations increase with the length of the RNA. Due to both the flexibility of long RNAs and the presence of alternative conformations, typical biophysical assays such as NMR and crystallography are only effective for a subset of small RNAs less than 150 nt[35]. Chemical probing techniques, in combination with next-generation sequencing, are a prominent method to study secondary structure of RNA. Chemical modifications are made to unpaired (open) nucleotides in an RNA, but base-paired nucleotides are unable to be modified. Analysis of the patterns of modification allows for prediction of secondary structures based on possible base-pairing combinations. Dimethyl sulfate (DMS) is a chemical probe used in the studies described in the following chapters.

HIV-1 makes use of several known RNA structures, and many more functionally important RNA structures are hypothesized. HIV-1 RNA structures can exist in an equilibrium or in mixtures of alternative structures. The virus can take advantage of the generation of ensembles of RNA structures in order to regulate biological processes. Some of the most prominent HIV-1 RNA structures and their roles in the replication cycle are highlighted in this section.

5'UTR

The HIV-1 RNA 5'UTR is a complex and highly structured region (Figure 1.2). The 5'UTR also is one of the most highly conserved sequences of the entire HIV-1 genome. It consists of a series

of functional RNA structures[36]. The first structure is TAR, which is an extremely stable single stem-loop. After TAR is a stem-loop that occludes the 5'-poly-adenosine tract (poly-A). The structure begins to become more complex, and no single model exists for the remaining features. A large RNA element that contains the primer binding site (PBS), complementary to the $tRNA^{Lys3}$, is found after the poly-A stem. Next are a series of four stem-loops, the first of which contains the major splice donor site and the third of which has the packaging signal (psi)[37]. The psi stem-loop contains the dimerization initiation site (DIS). DIS refers to a palindromic 6 nt sequence that is necessary but not sufficient for packaging and is only part of the minimum necessary psi[38]. Dimerization extends beyond just the 6 nt sequence once two genomic RNAs interact. Stabilization of the psi stem-loop, which increases the number of base-pairs to be broken for extended base-pairing between the dimers, decreases viral fitness[39].

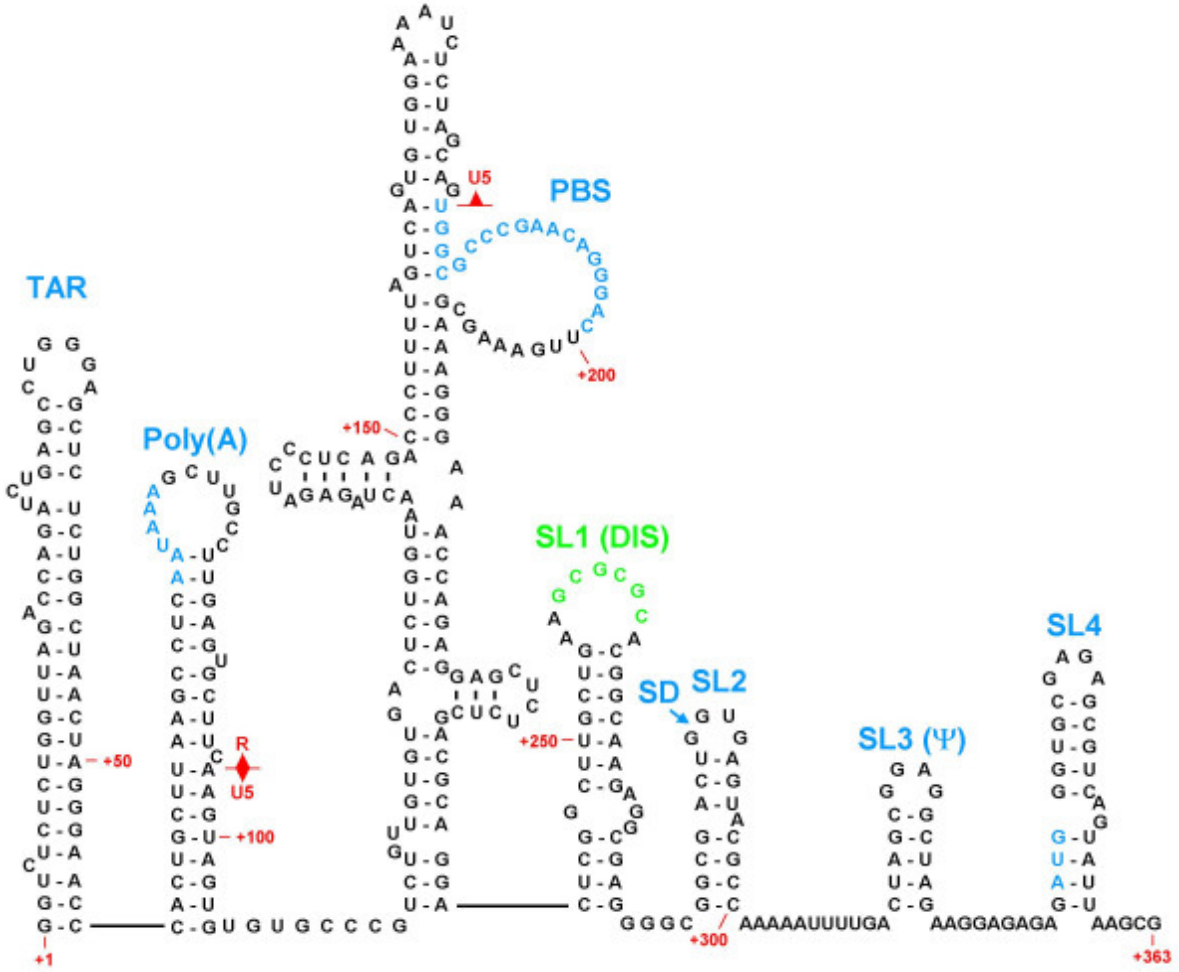

Figure 1.3*: Illustration of HIV-1 5'UTR. The sequence and numbering based on reference strain HXB2. Image reproduced from Russell et al.*[40]

The 5'UTR has been shown to form alternate conformations. A long-standing hypothesis is that there is a structure of the 5'UTR that inhibits translation and promotes dimerization and a structure that allows translation to occur. This observation was first made by *in vitro* transcribed and refolded 5'UTR running in two bands on a native gel. Mutations that stabilize or destabilize certain stem-loops could make the RNA run as a single band[41]. In-gel Selective 2'-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) confirmed that the two bands run as two distinct structures[42]. NMR was used to show that the two proposed structures form *in vitro* in

small sections[43,44]. This hypothesis gained more evidence recently with the discovery that the transcription start site of HIV-1 can include between one and three G's. Three G's were shown to promote the translatable, monomeric structure[45]. A recent Förster Resonance Energy Transfer (FRET) assay confirmed the existence of multiple conformations and found that binding thermodynamically stabilized the structure that supports dimerization[46]. Additionally, a careful study using NMR and polysome profiling showed a correlation between the number of base pairs in the stem-loop containing the Gag start codon and translation of Gag[47].

Despite several lines of evidence that the 5'UTR exists in multiple conformations, there is no unifying model of the structure. This is likely because each study uses different systems, truncations, strains and folding conditions. Also, individual elements of the 5'UTR can exist in multiple conformations. Even TAR was found to have a rare, transient alternate structure[48]. The alternate structure is unable to bind Tat. This alternate structure is hypothesized to be a negative feedback mechanism to ensure the Tat positive feedback loop does not progress too rapidly for virions to be assembled before the cell dies. How the alternate structures of each element fit into the overall model of 5'UTR structure and favor either the translated or packaged form is still the subject of research.

Rev Response Element

The Rev Response Element (RRE) is a highly structured region of the HIV-1 genome located within the Env coding region. The RRE binds to the viral protein Rev, and together they allow unspliced and incompletely spliced HIV-1 RNA to be exported from the nucleus[49-51]. Normally, a fully spliced RNA is exported from the nucleus by the mRNA transport pathway dependent on

Nxf1. The Rev-RRE interaction allows the HIV-1 RNA to be exported through an alternate pathway dependent on Crm1. MS HIV-1 RNA, which does not contain the RRE, is able to exported through the Nxf1 pathway.

The structure of the RRE has been studied intensively. The structure of HIV-1_{NL4-3}, a common lab strain, was found to exist as a mixture of two structures[52]. One structure consisted of a series of five stem-loops and the other had four stem-loops. Rev was shown to bind initially at the junction of stem-loops I and II. Initial binding of RRE causes a tertiary structure change that favors other molecules of Rev to multimerize across the entire RRE structure[53]. The structure of RRE is sensitive to mutations. The two-structure mixture from HIV-1_{NL4-3} is not necessarily found in all strains[54]. On a native gel, *in vitro* transcribed and refolded RRE sequences from different primary HIV-1 isolates all run in distinct patterns[55]. That result suggests that each RRE has a unique structure. The maximal nuclear export activity was observed when the Rev and RRE for each participant was matched. Similar results were obtained when matching RRE sequence and Rev from viruses obtained at different points over time, indicating that the structure of the RRE and Rev co-evolve within an infected individual[56].

Gag-pol frameshifting element

The HIV-1 Pol polyprotein, which consists of PR, RT and IN, is only translated by about ~5% of ribosome translation events. When Pol is translated, the ribosome slips on a heptanucleotide motif and translocated -1 in frame[32,33]. The slippage allows it to bypass a stop codon and continue translating the entirety of the Pol ORF. The Gag-Pol transition has extensive complementarity, creating a highly structured region that has been proposed to be either a series

of pseudoknots or double-stranded helices[57-59]. Mutations made to affect the stability of the RNA structure in this region reduce frameshifting and viral fitness, showing that frameshifting is regulated by RNA structure[60,61]. The Gag-Pol region of another retrovirus, Murine Leukemia Virus (MLV), has been shown to have alternative conformations of pseudoknots that produce an equilibrium of two structures. The predominate structure inhibits translation and the less abundant structure encourages slippage. The two structures exist in an equilibrium that exactly corresponds to the percentage of read-though translation that occurs[62]. For HIV-1, a more complex but similar model has been proposed. The Gag-Pol structure has been shown to be highly conserved and alternate conformations have been identified by chemical probing[57,58].

Other HIV-1 RNA structures

The entire structure of the HIV-1 genome was measure by SHAPE[59]. One novel finding was the identification of a highly stable stem at the signal peptide sequence of Env, which may slow down translation enough to ensure that the Env protein is properly embedded in the endoplasmic reticulum membrane. The structure of HIV-1 was compared to other lentiviruses using the same conditions, and 5 structures were highly conserved[63]. Interestingly, three structures occurred at the boundary of proteins in the same polyprotein (for example, at the PR-RT boundary in Pol). It was hypothesized that these structures slow translation and allow the protein to be folded properly before continuing to the next protein.

Many open questions remain regarding HIV-1 RNA structure. One key limitation of the studies described above is that they have been done largely using in vitro transcribed and refolded RNA. Experiments with full-length RNA in a cellular context are very rare in the field of RNA

structure. Determination of complex RNAs with alternate conformations in a cellular context have not been performed but would confirm the biological relevance of the hypothesized RNA structures. Given the abundance of structure across the genome and possibility of novel RNA structure regulatory elements, HIV-1 RNA structure remains an important open area of investigation.

1.4 HIV-1 splicing and splicing regulation

HIV-1 splices to produce over 100 unique splice products, although the majority of RNA remains unspliced (Figure 1.4)[31,64]. The unspliced RNA is ~10kb, the SS products are ~4 kb and the MS products are ~1.8 kb[30]. All SS products utilize the major splice donor site, or D1, located in the 5'UTR upstream of the Gag start codon. D1 is spliced to one of the splice acceptors A1-5. The splice acceptor that is used determines the translation product, as each of the splice sites lies just upstream of an HIV-1 ORF start codon. For A4 and A5, multiple splice acceptors in close proximity yield the same translation product. MS products utilize the D1:A1-5 junction as well, but also include a D4:A7 junction which removes most of the Env coding region, including the RRE. The D2 and D3 sites lie just downstream of A1 and A2. Small exons made of the nucleotides between A1:D2 and A2:D3 can be added to any transcripts that utilize A3-5, increasing the number of possible splice products. The function of these two exons is still the subject of research.

Many questions remain about the regulation of HIV-1 splicing. The current understanding of splice regulation involves the relative weakness of the splice acceptor sites and the balance of splice enhancer and silencer sites for RNA binding proteins (RBP)[65,66]. The strength of splice donor sites is determined by the complementarity to the 8 nt small nuclear RNA U1 recognition motif. The strength or weakness of splice acceptor sites is dictated by their affinity for U2AF, partially determined by the length and consistency of the polypyrimidine tract immediately upstream of the dinucleotide AG where the splicing occurs[66]. HIV-1 splice acceptor sites are generally weak compared to human splice acceptor sites, but their suboptimal activity is necessary for splice product regulation. Mutation of downstream splice acceptors or donors into

stronger splice sites, it decreases viral fitness and changes abundance of other splice products[67,68].

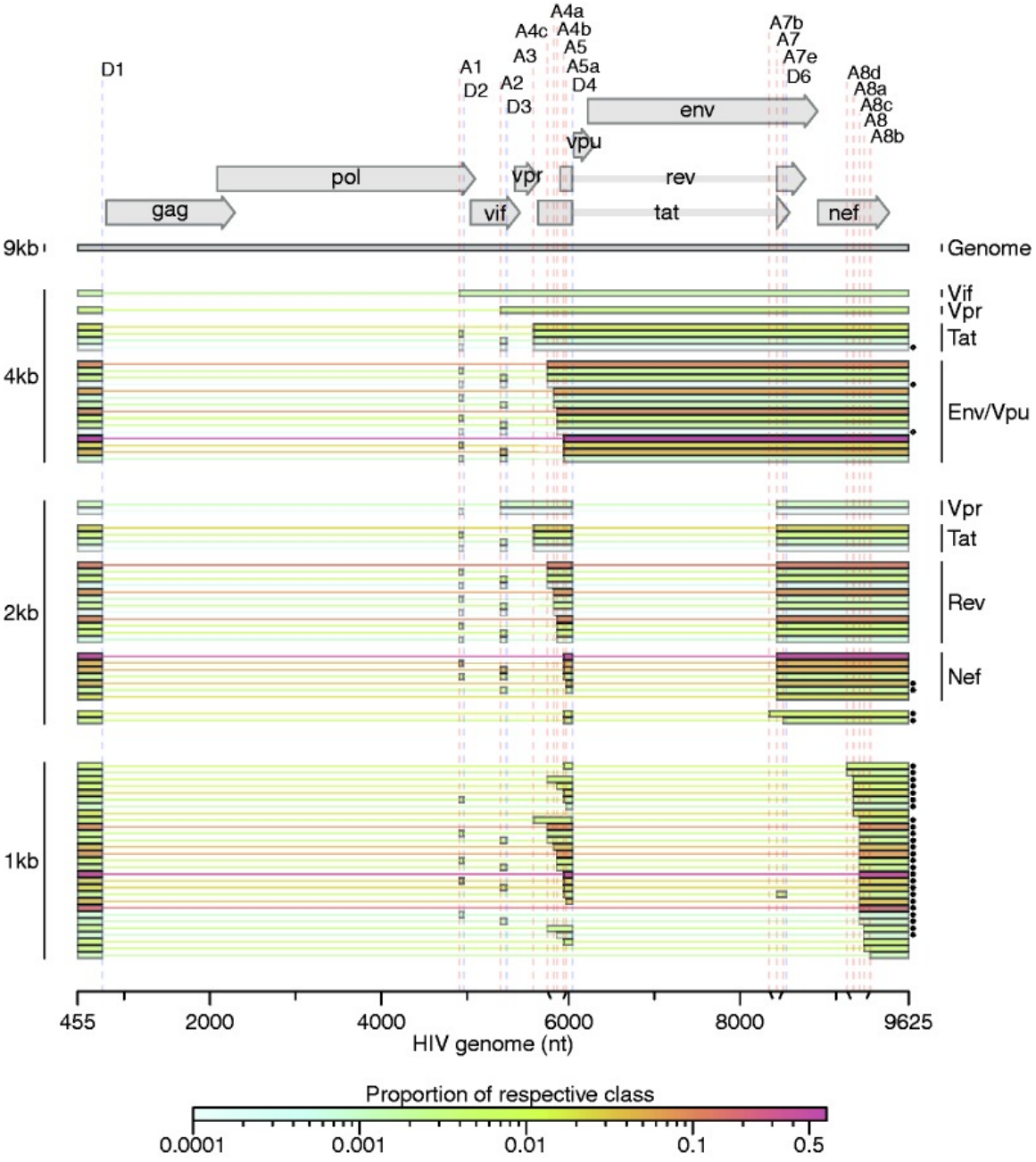

Figure 1.4*: Overview of HIV-1 splicing. On the top of the image is the genomic organization of the HIV-1 genome, along with the location of the splice donor and acceptor sites. Under that are the different size class of HIV-1 transcription products and unique splice products. Figure reproduced from Ocwieja et al.*[31]

The splice acceptor sites are influenced by the binding sites of mutually antagonistic RBPs from two families, the serine-rich arginine-rich proteins (SR proteins) which enhance weak splice site usage and the heterogeneous nuclear ribonucleoproteins (hnRNP), which inhibit splice acceptor usage. Each splice acceptor site has a number of both silencer and enhancer recognition sites in the immediately upstream or downstream sequence. Enhancer sites are located within exons (exonic splice enhancer-ESE), silencer sites are found in both exons and introns (exonic splice silencer-ESS and intronic splice silencer-ISS)[65,66]. The balance of binding of the antagonistic proteins sets the basal splice acceptor usage. Some SR and hnRNP binding sites in the HIV-1 overlap slightly with one another, and so they enhance or suppress splicing based on the expression levels of the human RBPs.

However, the current model of splice enhancer and inhibitor was made using mostly reporter assays and mini-genome systems. An intriguing recent study using a non-biased, deep-sequencing approach to study RBP binding in the context of cells infected with full-length virus showed that hnRNP A/B binding is more distributed throughout the HIV-1 genome than previously thought. The study also found that hnRNP H1 enhances splice acceptor usage[69]. These data challenge our current understanding, and show we cannot completely explain splicing regulation in HIV-1. Another open question is why so much of the HIV-1 remains unspliced when human RNAs get completely spliced, even with weak splice sites and the presence of splice silencer sites.

Another mechanism by which HIV-1 regulates splicing is by the use of Rev. Rev is hypothesized to regulate splicing by exporting the incompletely spliced RNA from the nucleus before the

spliceosome has time to fully splice the RNA. Mutants that enhance splice acceptor strength decrease the total Rev activity, as transcripts are spliced before they are able to be exported[70]. Interestingly, Rev activity is dependent on partial spliceosomal assembly at the D4 splice site, indicating that Rev may interact with the spliceosome to interact in turn with the RRE[71-73].

There is also some evidence that RNA structure regulates splicing. One study that provided evidence for RNA structure across the HIV-1 genome involved a series of synonymous mutations in sections across the HIV-1 genome[74]. The mutations had several phenotypes involving mis-splicing. Some series of mutations caused the virus to overuse splice acceptor sites. Some mutations caused cryptic splice donor sites to be exposed and used. These experiments provide evidence that RNA structure is involved in HIV-1 splicing regulation. Additional evidence to support this hypothesis is that culturing HIV-1 at higher and lower temperatures changes the splice acceptor usage patterns, particularly at A1 and A2[64]. RNA structure is especially susceptible to changes in temperature. The two splice sites that had this observed phenotype also had the strongest over splicing phenotype in the synonymous mutation study. The D1 splice site has also been shown to be susceptible to regulation by RNA structure. If the stem-loop in which this site is located is strengthened, splicing is reduced because the splice site is not able to be bound by U1[75]. This observation helps to explain why the D1 splice donor site is not utilized 100% of the time even though it has perfect complementarity to U1, theoretically making it a strong splice donor. RNA structure has also been shown to regulate binding of splicing regulatory factors. In contrast to the stem-loop the obfuscates D1, the regulatory element ESS3, which suppresses A7 usage, is exposed in the loop of a stem-loop.

Mutations that destabilize the stem-loop decrease binding of hnRNP A in an *in vitro* binding assay[76].

1.5 HIV-1 Latency

The steps of the HIV-1 replication cycle described above characterize productive infection. However, cells can also become latently infected. Latently infected cells theoretically transcribe and translate no HIV-1 RNA or proteins. The totality of latently infected cells in an individual living with HIV-1 is referred to as the latent reservoir. The stability of the latent reservoir is maintained by the longevity of the latently infected cells and their capacity for homeostatic proliferation without reactivating the provirus. The latent reservoir persists for decades of ART[77,78]. Due to the stability of the latent reservoir, individuals living with HIV-1 must remain on life-long ART in order to maintain viral suppression[79]. If ART is stopped, 95% of infected individuals have sustained viral rebound within 8 weeks[80]. The reservoir is seeded very early in infection[81]. Individuals treated within the first stages of HIV-1 infection still rebound even after years of viral suppression[82]. Ultimately, the purpose of studying HIV-1 latency is to be able to find strategies to eliminate the reservoir from infected individuals.

Identity of latently infected cells

In a landmark study, central memory $CD4^+$ T (T_{CM}) cells and transitional memory $CD4^+$ T (T_{TM}) cells were found to be the most likely T cell subsets to have HIV-1 proviral DNA after prolonged ART[83]. T_{CM} cells are the stable, memory CD4+ T cells that reside mostly in lymph nodes but traffic throughout the body to surveil for their cognate antigen. They are capable of activating in response to antigen presented along with a co-stimulatory marker. T_{CM} cells mature into effector memory cells, or T_{EM} cells, which are not as long-lived and are capable of homing to inflamed tissues, rapid expansion upon binding of IL-15 and releasing cytokines. T_{TM} cells have an intermediate phenotype between T_{CM} and T_{EM} cells[84].

When measuring replication-competent virus by quantitative viral outgrowth assay, T_{CM} cells were found to have a much larger percentage of the reservoir than T_{TM} cells. T_{CM} cells were the largest component of the reservoir regardless of whether the infected individuals initiated treatment in the acute phase of infection or the chronic phase[85]. Other $CD4^+$ T cell subsets have also recently been shown to contribute more than initially thought to the latent reservoir. A small subset of memory $CD4^+$ T cell with high homeostatic proliferative capacity, T stem cells (T_{SCM}), were shown to contribute more to the reservoir over time. Although T_{SCM} cells constitute a small proportion of the reservoir at first, their heightened ability to maintain themselves leads to a larger contribution longitudinally[86]. Naïve $CD4^+$ T cells (T_N) also contribute more to the reservoir than previously thought. T_N cells were recently shown to release as many virions per million cells as T_{CM} cells. This finding implies that the rarer latently infected T_N cells are more likely to release large quantities of virus than T_{CM} cells[87].

Mechanisms of HIV-1 latency

Most research has focused on the mechanisms of transcriptional regulation of the HIV-1 provirus. After integration, the provirus becomes associated with multiple histones to form nucleosomes, the most important of which are located near the promoter region of the provirus[88]. Epigenetic modifications made to the nucleosome at the promoter of HIV-1 can lead to the inhibition of transcription. Histone acetylation generally favors transcription and methylation inhibits transcription. Therefore, histone deacetylases and methyltransferases are important factors for the repression of HIV-1 transcription[88-92]. Even if the promoter is accessible, key transcription factors, in particular NF-κB, are sequestered in the cytosol while the T cell is in a

resting state[93]. Furthermore, the levels of RNA polymerase II elongation factor PTEF-b are reduced in resting CD4 T cells, preventing proper elongation of transcription even if initiation is successful[94]. The three main mechanisms of transcriptional repression described above are all reversed in the course of natural HIV-1 reactivation when the $CD4^+$ T cell goes from a resting to an activated state, and T cell activation is negatively correlated with T cells' ability to become latently infected[95].

In addition to transcriptional regulation, recent research finds that post-transcriptional barriers to reactivation exist as well. Using a carefully designed series of PCR reactions to capture key portions of HIV-1 RNA, it was found that HIV-1 is regulated at elongation, poly-adenylation and splicing[96]. Nuclear export of MS HIV-1 RNA is also thought to be restricted in resting $CD4^+$ T cells compared to activated $CD4^+$ T cells[97]. Export and translation of MS RNA must happen before US or SS RNA can be exported, as Rev is a MS RNA product.

Clonal expansion of the latent reservoir

A key question of HIV-1 latency has been how the reservoir is maintained over decades on ART. Clonal expansion of latently infected memory $CD4^+$ T cells is increasingly appreciated as a crucial contributor to reservoir maintenance. Clonal expansion refers to the proliferation of a latently infected $CD4^+$ T cell which leads to the progeny containing an integrated provirus. The hallmarks of a clonally expanded provirus are that all the clones have the exact same sequence and the exact same integration site. Clonal expansion was hypothesized long before it was proven, as deep sequencing needs to be done simultaneously for both full-length HIV-1 genomes and integration sites. The first time this was formally shown, there was an overrepresentation of

integration sites near genes involved in the cell cycle, indicating that the integration site may impact the ability of the host cell to expand[98,99]. Subsequent research on clonal expansion shows that the clones expand and contract over time[100]. They are also subject to immune pressure, and so over time while the level of HIV-1 DNA may appear constant, the clones that either have defects or are more silenced are selected for by the immune system[101]. Therefore, the reservoir is decaying by a higher rate than previously estimated, although it is still a slow decay[102]. Intact proviruses that are integrated into heterochromatin or other non-transcriptionally active regions of the genome also accumulate over time, given their deeper state of latency[103]. Clonally expanded sequences can be detected even if an infected individual is treated early[104]. Clones constitute a larger percentage of the replication-competent reservoir over time, indicating that clonal expansion is an important factor to consider when trying to disrupt latency[105].

Defective Proviruses

The true size of the replication competent reservoir is obscured by the presence of cells with integrated proviruses that are defective and cannot produce infectious virions. Defective proviruses are generated by a variety of mechanisms. One of the common mechanisms is hypermutation by a member of the human protein family apolipoprotein B mRNA editing enzyme, catalytic peptide-like (APOBEC). The APOBEC proteins are a family of intrinsic cell defense factors upregulated during viral infection. When an infected cell expressed APOBEC, the protein can be encapsulated in the budding virions. Inside the virion, the enzyme removes the amine group of cytosines of the negative, single-stranded viral DNA during reverse transcription[106]. Each member of the family has slightly different sequence specificity, and for HIV-1 the most important is APOBEC3G; APOBEC3F acts on HIV-1 to a lesser extent[107].

When the positive strand is generated based on the modified negative strand, the HIV-1 genome has an abundance of G->A mutations, which result in premature stop codons in the ORFs. Although the provirus can be integrated, it cannot make infectious virus. However, it is possible for some hypermutants to transcribe HIV-1 RNA and even translate some HIV-1 peptides. Antigen from reactivated hypermutated proviruses is hypothesized to exhaust and confound the immune response during ART-suppressed HIV-1 infection[108].

HIV-1 reverse transcriptase itself can be responsible for the defect that renders a provirus replication incompetent. As mentioned before, the two strand transfer events provide an opportunity for large internal deletions and inversions[7,8]. Small deletions at crucial viral elements, frameshift mutations, or nonsense mutations can also lead to defective provirus. Depending on the nature of the deletion or mutation, the provirus may be able to produce transcripts, protein and antigen[108].

Latency Reversal

One approach to eliminate the latent reservoir is to treat the infected individual with a drug that is meant to transcriptionally activate the HIV-1 provirus, which should theoretically lead to HIV-1 translation and antigen production[109,110]. Drugs that reactivate latently infected cells are referred to as latency reversing agents, or LRAs. The immune system can then identify and kill all the latently infected cells. LRAs are administered along with continued ART in order to prevent the reactivated virus from spreading. This approach is referred to as 'shock and kill'[110,111]. The natural stimulus for reactivation is T cell activation, whether through T cell receptor (TCR) activation or by cytokine stimulation. Therefore, it is logical to recapitulate this

pathway in order to induce reactivation. However, global T cell activation is not clinically viable due to immunopathological effects. Several factors that interact with components of the TCR pathway, most notably protein kinase C (PKC) agonists, have been evaluated *in vitro* and in clinical trials for HIV-1 latency reversal[112-115]. Bryostatin-1, a PKC agonist, has been used safely in clinical trials however the concentration was not high enough to have an effect[116]. Cytokines that do not fully activate T cells, but promote proliferation or cell state change have been investigated as well. IL-15 and IL-15 super agonist are other leading candidates for latency reversal[117,118]. IL-15 has also been shown to sensitize infected $CD4^+$ T cell to killing by cytotoxic $CD8^+$ T cells[118,119]. IL-15 promotes proliferation and survival of $CD4^+$ T cells, and the effect on activation is dependent on concurrent TCR stimulation. IL-7, which only promotes homeostatic proliferation but has no effect on activation status, was shown to increase the size of the latent reservoir. Latently infected cells are able to expand without reactivating or getting detected by the immune system[120]. Toll-like receptor (TLR) 9 agonist has been investigated to simultaneously induce HIV-1 transcription and immune surveillance of infected cells through natural killer and dendritic cells[121-123].

Aside from recapitulating T cell activation, another strategy for latency reactivation is to target some specific mechanism of transcriptional repression of HIV-1. There are several mechanisms of transcriptional restriction which have been targeted for HIV-1 latency reversal. Epigenetic silencing is a clinically relevant target, especially by histone deacetylases inhibitors (HDACi). Romidepsin, vorinostat and panobinostat have all been used clinically and have shown at best modest increases in HIV-1 RNA expression. However, no decrease in the viral reservoir measured by HIV-1 DNA or viral outgrowth has been observed in any trial[110,124-126].

Methyltransferase inhibitors have also been evaluated *in vitro*, however there is less clinical experience with these agents, so they have not advanced to clinical trials[127]. Disulfiram, a phosphatase PTEN inhibitor used to treat alcoholism, has also been tested in clinical trials. Similar to HDACi, dilsulfiram treatment achieved a modest increase in HIV-1 RNA, but no decrease in the reservoir[128,129]. Bromodomain inhibitors, namely JQ1, are meant to inhibit the non-coding RNA 7SK form associating with PTEF-b[130,131]. However, 7SK is not the main regulator of PTEF-b in primary cells[132]. As no one agent has so far been successful, combinations of LRAs are also being investigated. LRAs with distinct mechanisms of action have been shown to synergize[133]. Many clinical trials of LRA combinations, most frequently involving an HDACi and an immunomodulatory agent, are ongoing[134].

1.6 Conclusions

Adapting next-generation sequencing techniques to study HIV-1 biology requires careful consideration of HIV-1 biology. For RNAseq and RNAseq derivatives, one of the most difficult variables is the unique splicing pattern of HIV-1. Leveraging these new techniques is essential to answer long-standing questions about how HIV-1 regulates its replication cycle and also to generate novel hypotheses to answer the most important questions in the HIV-1 field today. One of the most pressing areas of research is HIV-1 latency, with a renewed focus on the basic biology of how transcription is regulated from the provirus.

In the next chapters, I will discuss the development of DMS-MaPseq for studying RNA structure of HIV-1 in virions and in cells. The development of this technique, in combination with a novel algorithm to detect alternative structures from chemical probing data, was used to identify novel HIV-1 RNA structures and probe known structures in cells. We discovered novel HIV-1 RNA structures that regulate splicing. A separate technique, CaptureSeq, was developed to study HIV-1 and human transcription after treatment with latency reversing agents. The enrichment allows for an in depth and sensitive analysis of HIV-1 RNA species that previous studies were unable to find.

Chapter 2: Viral RNA structure analysis using DMS-MaPseq

Phillip Tomezsko[1,2,3], Harish Swaminathan[3] and Silvi Rouskin[3*]

[1]*Program in Virology, Harvard Medical School, Boston, Massachusetts, USA*

[2]*Brigham and Women's Hospital, Boston, Massachusetts, USA*

[3]*Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, USA*

P.J.T. and S.R. developed the concept. P.J.T. performed the experiments. P.J.T., H.S and S.R. analyzed the data. H.S. wrote the publicly accessible analysis software. P.J.T. wrote the manuscript. P.J.T., H.S and S.R. revised the manuscript.

This work was accepted by *Methods*, in press.

Abstract

HIV-1 uses RNA structure to regulate key steps in its replication cycle. RNA structure is difficult to study due to the numerous conformations a given RNA sequence can form and the flexibility of the structure. The molecular environment, especially in a cell with various RNA binding proteins expressed, can also impact how RNA folds. Chemical probing, which add modifications to RNA in order to differentiate open and closed bases, is a leading method for RNA structure analysis. Combined with deep-sequencing, chemical probing can resolve RNA structures in a high-throughput manner.

DMS-mutational profiling and sequencing (MaPseq) is a powerful chemical probing and deep-sequencing method to analyze RNA structure in cells, in virions and of *in vitro* refolded RNA. Among other chemical probing methods, DMS-MaPseq stands out as highly robust and adaptable to different conditions and cellular systems. DMS-MaPseq improves the existing DMS-seq strategy by incorporating multiple DMS modifications per sequencing read. This feature allows for more information for each base, enabling structure determination of genes expressed at a low abundance. Compared to purely computational approaches, DMS-MaPseq is able to account for the impact of cellular factors and other considerations that are unable to be included in models. Chemical probing techniques bypass the sequence length limitations of classic structural methods such as X-ray crystallography and NMR.

Mutational profiling and sequencing approaches have been used over the last couple years to answer key questions about viral RNA. In this chapter, I will discuss our efforts to adapt DMS-MaPseq for HIV-1 RNA structure analysis. The key troubleshooting steps were titration of DMS

concentration and length of incubation with DMS during modification. The most crucial determinate for high-quality data was found to be the average number of DMS-modifications per sequencing read. The detailed development of DMS-MaPseq for HIV-1 built an experimental framework for further studies.

2.1 Introduction

Single-stranded RNA is able to form complex structures through both canonical and non-canonical base-pairing interactions and base-stacking[135]. The range of functions ascribed to RNA structure has expanded rapidly[136,137], precipitated by the discovery that the catalytic element of the ribosome is RNA[138-140]. All major classes of RNA, including mRNA, form structures that have functional importance[136]. RNA is able to form multiple alternative structures based on thermodynamic properties, but the structure is also influenced by the cellular environment, particularly by RNA binding proteins and RNA helicases[141,142]. Given these factors, prediction of biologically relevant RNA structures is extremely difficult by thermodynamic modeling alone, although there have been advances in algorithms[143] and methods to ensure biological relevance[144]. Several approaches exist to experimentally study RNA structure, including chemical probing. The basis of chemical probing is to add modifications to open RNA bases but not paired RNA bases, which can be used as constraints to greatly improve the accuracy of prediction programs[145]. Dimethyl sulfate (DMS), the chemical described in the methodology below, is the oldest and one of the most widely used chemicals for RNA structure probing[146]. DMS adds methyl groups to the N1 and N3 positions of unpaired adenosine and cytosine nucleotides (Figure 2.1). The first RNA structure assays converted the modifications into signal by comparing intensity of bands after poly-acrylamide gel electrophoresis of truncated reverse transcription products; in the case of DMS this was termed DMS footprinting[147,148].

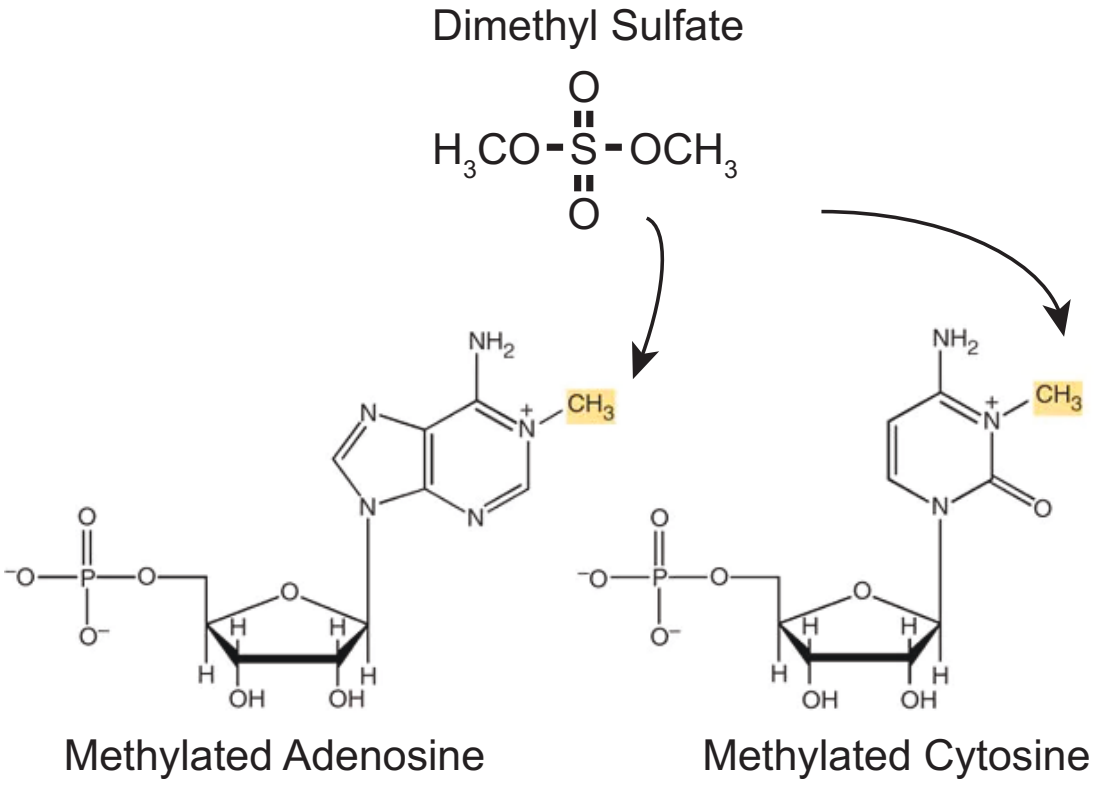

Figure 2.1*: DMS modification of adenosine and cytosine. The added methyl groups are highlighted in yellow. Figure adapted from Tijerina et al.*[148]

Advances in high-throughput sequencing allowed for DMS sequencing (DMS-Seq), the sequencing of reverse transcription termination products on a transcriptome-wide scale[141]. Discovery of reverse transcriptases that add random mutations when encountering a methylation, including thermostable group II intron reverse transcriptase (TGIRT-III), has allowed for development of DMS-mutational profiling and sequencing (MaPseq)[149,150]. In this technique, each read has multiple DMS-induced mutations, maximizing the information given per read (Figure 2.2). DMS-MaPseq provides single-molecule RNA structure information that can reveal

heterogenous RNA structures and maximize sequencing depth for low-abundance RNA species[150].

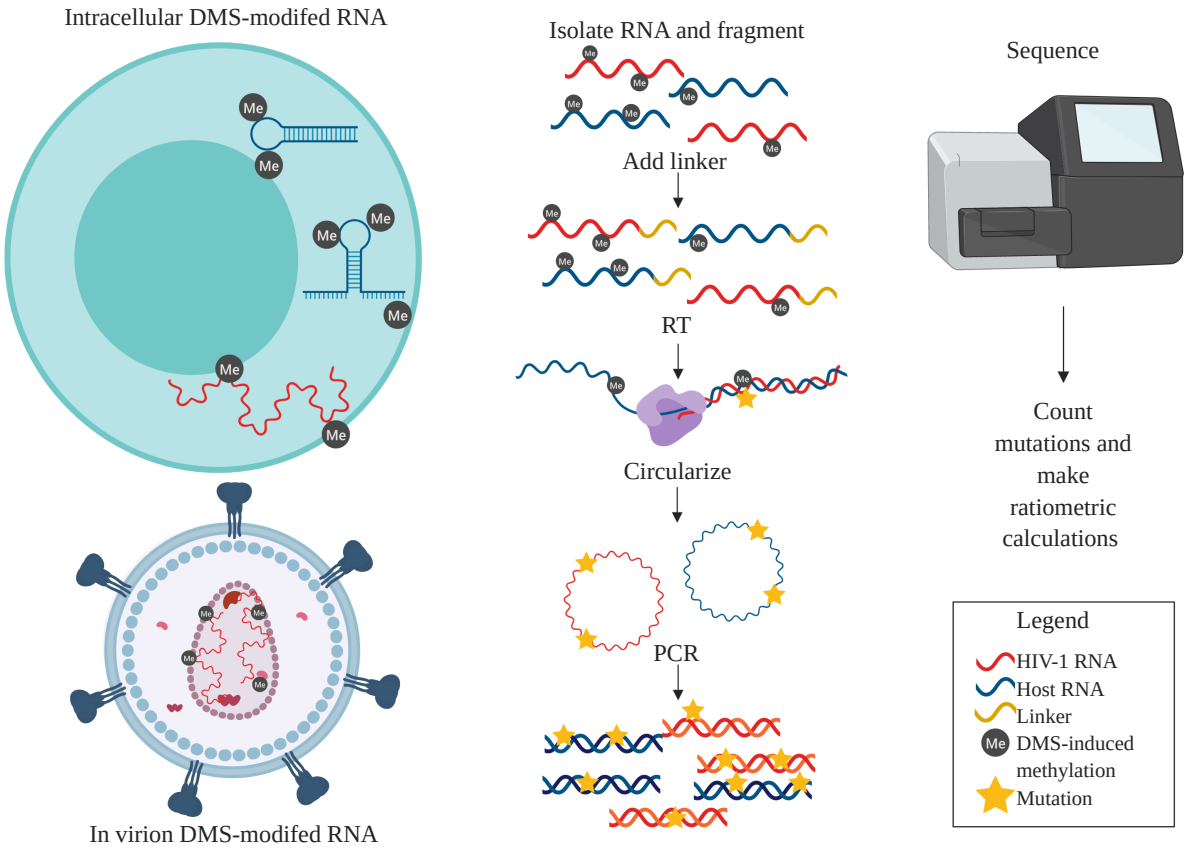

***Figure 2.2**: Overview of DMS-MaPseq method with infected/transfected cells and virions. The first column depicted DMS modification of RNA intracellularly or in virion. The second column shows the general library generation protocol. The third column shows sequencing and conceptual analysis.*

DMS-MaPseq has a number of advantages over other methods to investigate RNA structure, but it provides complementary information to these methods as well. X-ray crystallography and nuclear magnetic resonance (NMR) are regarded as the gold standards of RNA structure analysis; however, both methods are limited to *in vitro* folded RNAs. For X-ray crystallography and to a lesser extent NMR, the RNA must have an extremely rigid and homogenous structure in

order to provide interpretable signal. For NMR in particular, the RNA must be small (<155 nt) in order to be resolved[35]. One advantage compared to DMS-MaPseq is that X-ray crystallography and NMR provide 3D structural information. Cryogenic electron microscopy (cryoEM) is another extremely powerful tool to determine RNA structure *in vitro*. This method involves freezing purified RNA or RNA-protein complexes in a thin layer of ice and measuring the structure by electron microscopy. The images of the molecule in different orientations need to be sorted based on similarity in order to get a 3D projection of the target. Viral RNA-protein complexes such as the IRES of CrPV bound to the ribosome have unlocked insight into the mechanism of interaction[151]. More recent studies have used cryoEM to study the interaction of viral genomes and viral proteins, such as the nucleocapsid coating of the Hantaan virus and Ebola[152,153]. Although crystallization is more efficient in the presence of RNA binding proteins, single-stranded RNA and even naked viral genomes have been analyzed with this method[154]. When comparing cryoEM to SHAPE, the secondary structure of naked STMV genome structure was consistent between methods however there was significantly more heterogeneity identified by the cryoEM study[155,156].

Another commonly used chemical probing technique is called selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE)[157]. SHAPE and SHAPE-MaP use a family of acylating electrophiles to modify predominantly unpaired nucleotides in an RNA chain[158-160]. DMS-MaPseq and SHAPE-MaP share similar theory but one key distinction is that DMS-based techniques are far less sensitive to RNA-binding proteins as compared to SHAPE-based techniques[160]. This difference is due to the fact that DMS is smaller than SHAPE reagents and modifies the Watson-Crick face whereas SHAPE modifies the backbone. In order for a protein to

provide protection from DMS it must directly interact with N1 of adenosine or N3 of cytosine whereas many different types of RNA-protein interactions provide protection from SHAPE[160]. In a comparison of *in vitro* and *in vivo* structure therefore, DMS-MaPseq makes it possible to differentiate between the actual binding of the RNA binding protein and any structural change of the RNA as a result of the RNA binding protein.

RNA viruses, including HIV-1, have been an area of particular interest for the RNA field because of an abundance of well-studied functional RNA structures. Embedding function in RNA structure allows the virus to maximize the information contained in a small genome. Viral RNA structures are utilized in many diverse stages of the replication cycle, including packaging, replication, transcriptional regulation, translational regulation and anti-host defense[161]. A notable family of viral RNA structures is the internal ribosomal entry site (IRES) family, which were first discovered in picornaviruses[162]. IRESs allow the viral RNA to bypass some or all of the normally essential host translational cofactors associated with the cap-dependent translation and have been found in many RNA viruses[163]. RNA structures are also relevant to DNA viruses and have similar functions. HHV8 was the first DNA virus to be shown to use an IRES[164] but other IRES elements have been identified in other DNA viral transcripts, such as polyomavirus SV40[165]. Additionally, several classes of highly structured viral non-coding RNAs in gammaherpesviruses (such as HHV8 and EBV) contribute to transformation of the cell[166]. Even for the most well-characterized of viral RNA structures, many open questions remain that can benefit from a high-throughput, robust *in vivo* chemical probing methods such as DMS-MaPseq, including: 1) How do different cellular environments, most notably different expression patterns of RNA binding proteins, change the RNA structure? 2) How does the RNA structure change

during the course of infection? 3) How do alternative forms of the structure impact regulation of the structure's function? 4) How do mutations or natural viral sequence diversity change RNA structures?

DMS-MaPseq and SHAPE-MaP have been used to answer some of these questions for viral RNA structures. DMS-MaPseq was used to confirm the NMR structural prediction of an enterovirus IRES structure[167]. SHAPE-MaP was recently used to determine the differential binding of RNA binding proteins in the nucleus, cytoplasm and virion for the PAN RNA of HHV8, which is crucial for suppression of the host anti-viral response during infection[168]. Two intriguing reports on the structure of the influenza virus genome in cells using DMS-MaPseq[169] and in virions using SHAPE-MaP[170] have shed insight into how the viral segments form a complex network of structures and intramolecular interactions that allow for proper packaging and reassortment. Importantly, these structures are stable even in the presence of RNA binding proteins.

SHAPE-MaP has also been used to study the HIV-1 RRE in different contexts. The RRE is a stable structure that enables export of unspliced and incompletely spliced HIV-1 RNA from the nucleus of infected cells through interaction with the viral protein Rev[51]. Recent studies using SHAPE-MaP were able to uncover the binding site of two new inhibitors on the RRE[171,172]. Another study using SHAPE-MaP was able to track slight changes in structure of the RRE over the course of infection and correlate structural changes to activity of the RRE, thus showing that the RRE is under selection pressure to maintain or increase activity as HIV-1 evolves in the

host[56]. Here we present a detailed method for DMS-MaPseq of cells infected or transfected with HIV-1 in order to probe intracellular RNA structure.

2.2 Methods

HEK293t culture and transfection

HEK293t (ATCC) were cultured in Dulbecco's Modified Eagle Medium supplemented with 10% fetal bovine serum and Penicillin/Streptomycin (50 U/mL; ThermoFisher Scientific) at a concentration of ~1-2 million cells/mL during maintenance. For experimental setup, 0.9 million HEK293t cells were seeded per well of a 6-well plate and allowed to attach overnight. A plasmid containing HIV-1_{NHG} was transfected into the HEK293t cells using X-tremegene9 (MilliporeSigma) according to manufacturer's instructions. Cells were incubated for 48 hours in order to achieve peak virion production. This method also works with suspension cells grown at a density of ~1-2 million cells/mL before DMS-modification.

DMS-modification of HIV-1 virions

First, virions were isolated from the supernatant. All supernatant was removed from the plate and fresh medium was added. The supernatant was filtered through a syringe driven 0.22 μm filter (MilliporeSigma). The supernatant was centrifuged for 1 hour at 28,000xg, 4ºC. Supernatant was removed and the pellet was resuspended in 100 μL of ice-cold PBS. 100 μL of 2x modification buffer (400 mM NaCl and 6 mM $MgCl_2$) was added to virions and incubated at 37ºC for 10 minutes on a Thermomixer. 20 μL of DMS (MilliporeSigma) was added and incubated 10 minutes at 37ºC while shaking at 1000 rpm on a Thermomixer. 440 μL of β-mercaptoethanol (BME; MilliporeSigma) was added to neutralize the DMS. The RNA was purified with the Clean and Concentrator -5 kit (Zymo Research) according to manufacturer's specifications for all RNA fragments. The RNA typically was eluted into 30 μL of nuclease-free water, resulting in an RNA concentration of 20-50 ng/μL.

No difference was found when this protocol was tried with and without detergent (1% SDS) in the modification buffer to lyse the virion. We speculate the acidification of the medium from DMS disrupts the viral membrane and capsid.

DMS-modification of HIV-1+ HEK293t cells

The medium was pre-warmed to 37ºC before DMS-modification and 200 µL of DMS was added to 15 mL of warm medium (1.33% DMS final concentration). A range of DMS percentages (0.5-2.5%) were tried in the course of developing this protocol. Supernatant was removed from the wells of the plate, the cells were washed with PBS, and 2 mL of the DMS-media was added to each well. The plates were placed immediately in the incubator and incubated for 4 minutes. The medium was removed and 2 mL of ice-cold 1:1 PBS:BME was added to each well to neutralize the DMS. The cells were scraped off with a cell scraper, transferred to a 50 mL conical tube, and pelleted by centrifugation for 5 min at 1000xg, 4ºC. Supernatant was removed, the pellet was resuspended in 15 mL of ice-cold PBS and centrifuged for 5 min at 1000xg, 4ºC. Supernatant was removed and the pellet was resuspended in 1 mL of Trizol reagent (ThermoFisher). RNA was extracted according to manufacturer's specifications. The purified RNA was resuspended in 50 µL of nuclease-free water, and typically obtained RNA concentrations of ~2 µg/µL.

rRNA subtraction

As DMS-modified RNA is often fragmented, and the poly-A tail is susceptible to methylation that could interfere with Oligo(dt) beads from getting quality mRNA isolation, rRNA subtraction was necessary. Several commercially available rRNA subtraction kits including RiboZero

(Illumina; discontinued), FastSelect (Qiagen) and RiboMinus (ThermoFisher Scientific) were tested, each according to manufacturer's specifications. However, for cost considerations, an in-house oligo cocktail designed as described by Adiconis *et al*.[173] was also tested. This in-house method used 1-3 μg of RNA in 7 μL of nuclease-free water (it is possible to use multiple reactions for library generation and combine after rRNA subtraction) to which 2 μL of 5x hyb buffer (1M NaCl and 500 mM Tris-HCl pH 7.5) and 1 μL of rRNA subtraction mix were added. The reaction was run on a PCR machine -1ºC/minute starting at 68ºC and ending at 45ºC. Once the reaction reached 45ºC, 33 μL of nuclease-free water, 5 μL of Hybridase buffer and 2 μL of Hybridase Thermostable RNase H (Lucigen) were added. The reaction was incubated at 45ºC for 30 minutes. RNA was purified using the RNA Clean and Concentrator -5 kit and eluted into 44 μL of nuclease-free water. If multiple reactions were to be combined, elution volumes were adjusted to yield a final total volume of 44 μL. Five μL of 10x Turbo DNase buffer and 1 μLTurbo of DNase (ThermoFisher Scientific) were added to the eluted RNA, and the samples were incubated for 30 minutes at 37ºC. Turbo DNase Inactivation reagent (5.5 μL) was added and the samples were incubated for 5 minutes at room temperature. The samples were briefly centrifuged and the supernatants transferred to new tubes. The RNA was purified using the RNA Clean and Concentrator -5 kit and eluted into 9 μL of nuclease-free water. The purified RNA at this point can be used for either library generation or RT-PCR.

Library Generation

First, the RNA was fragmented. The samples were incubated at 95ºC for 1 minute in order to denature the RNA. One μL of 10x RNA fragmentation reagent (ThermoFisher Scientific) was added and the fragmentation reactions were incubated at 70ºC for 45 seconds. The samples were

placed on ice and 1 μL of 10x Stop solution was added. The RNA was purified using RNA Clean and Concentrator -5 following instructions for all fragment size collection and eluted in 6.5 μL of nuclease-free water.

The next step was to dephosphorylate the RNA fragment ends and add a linker to the 3' end. To 6.5 μL of each sample, 1 μL of CutSmart buffer, 1 μL of Shrimp Alkaline Phosphatase (New England Biolabs) and 1 μL of RNaseOUT (ThermoFisher Scientific) were added. The reactions were incubated at 37°C for 1 hour. Subsequently, 6 μL of 50% PEG 8000, 2.1 μL of 10x T4 RNA Ligase Buffer, 2 μL of T4 RNA ligase 2, truncKO (ThermoFisher Scientific) and 1 μL of 20 μM N12 linker were added and the reactions incubated for 18 hours at 22°C. RNA was purified using RNA Clean and Concentrator -5 kit for all RNA fragments and eluted into 15 μL of nuclease-free water. Excess linker was degraded by addition of 2 μL of RecJ buffer, 1 μL of RecJ exonuclease (Lucigen), 1 μL of 5' Deadenylase (New England Biolabs) and 1 μL of RNaseOUT. The samples were incubated for 1 hour at 30°C. The RNA was purified using Clean and Concentrator -5 kit, following directions for RNA >200 nt. By only purifying large fragments, more of the excess linker was removed. The samples were eluted into 11 μL of nuclease-free water.

For reverse transcription, the following reagents were added to the samples:

- 4 μL of M-MLV reverse transcriptase buffer (ThermoFisher Scientific)
- 1 μL of dNTP mix
- 1 μL of 0.1M DTT
- 1 μL of 10 μM Library RT Primer

- 1 μL of RNaseOUT
- 1 μL of TGIRT-III (Ingex)

Reactions were incubated at 65ºC for 1.5 hours, after which 1 μL of 4 N NaOH was added and the samples were incubated at 95ºC for 3 minutes to degrade RNA. Twenty μL of 2x TBE-Urea sample loading buffer (ThermoFisher Scientific) was added and the samples were loaded onto a 10% TBE-Urea Novex gel (ThermoFisher Scientific). The gels were run for ~2 hours at 180V. The gels were stained with SybrGold (ThermoFisher Scientific). The bands of the expected size were excised on a blue light box. The gel fragments were extruded through a punctured 0.65 mL Eppendorf tube and collected in a 1.5 mL Eppendorf tube. The DNA was extracted from the gels by adding 400 μL of 300 nM NaCl to each sample followed by incubation while shaking at 70ºC. The samples were placed in a 0.22 μm Costar Spin-X column (MilliporeSigma) and centrifuged at maximum speed for 30 seconds, after which the columns were discarded. 500 μL of 2-propanol and 3 μL of glycoblue (ThermoFisher Scientific) were added and the samples were frozen on dry ice. The samples were centrifuged for 45 minutes at 18,000xg, 4ºC. The supernatant was removed and the pellets were washed with 250 μL of ice-cold 70% ethanol. The pellets were resuspended in 15 μL of nuclease-free water. To each sample, the following reagents were added:

- 2 μL of 10x CircLigase reaction buffer
- 1 μL of 1 mM ATP
- 1 μL of 50 mM $MnCl_2$
- 1 μL of CircLigase (Lucigen)

The reactions were incubated for 2 hours at 60ºC and then 10 minutes at 80ºC. In a fresh PCR strip, the following reagents were added:

- 11 μL of nuclease-free water
- 4 μL of 5x HF Phusion Buffer
- 0.5 μL of Phusion (New England Biolabs)
- 1 μL of 10 μM library reverse primer
- 1 μL of 10 μM library forward primer
- 0.5 μL of dNTP
- 2 μL of circularized cDNA

The following PCR program was run:

i. 30 seconds at 95°C
ii. 15 seconds at 95°C
iii. 5 seconds at 55°C
iv. 10 seconds at 65°C
v. Go to ii for 8-14 cycles
vi. Hold at 4°C

The PCRs typically were run for 10 and 12 cycles on the first attempt, and the number of cycles adjusted if the PCR needed to be rerun. After PCR, 2 μL of 6x loading dye (ThermoFisher Scientific) was added to the samples, which were then electrophoresed through an 8% TBE gel for ~50 minutes at 180V. The gels were stained with SybrGold. The gel extraction process was repeated from gel extrusion to 2-propanol precipitation and centrifugation. After the 70% EtOH wash, the DNA was resuspended in 11 μL of nuclease-free water. One μL of the sample was diluted 1:5 and submitted to Bioanalyzer in order to check the concentration and size of the DNA product.

Sequencing

Although longer-sequencing reads are generally preferred, due to the fragmentation induced by DMS, there was a need to balance optimal read-length. We had success running 75x75nt, 150x150nt and 300x300nt on Illumina iSeq, MISeq and HIseq respectively. The right sequencing length also relied on the target insert size. A target insert size of ~100-250 nt was optimal, depending on the desired sequencing length.

Quality Control

FastQC was run on all samples in order to summarize the basic quality of the reads. Next Trimgalore was used to remove bases of reads with a Phred score <20 from the reads. The '--fastqc' option was included to run FastQC on the post-trimming reads.

Mapping

Sequences were aligned with Bowtie2, which also provided a transcriptome or genome for the reference as needed[174]. Alternatively, a splice-aware aligner such as HISAT2 was used[175]. For alignments performed with Bowtie2, the following options were used: '--local --no-unal --no-discordant --no-mixed -X 1000 -L 12'. This set of options allowed for the maximum number of mismatches by reducing the seed length and running with the local option. Allowing for short seed length was important since mismatches occurred throughout the reads because of the DMS-induced mutations. The remaining options reduced run time and prevented other technical artifacts stemming from discordant and mixed pairs. The output of Bowtie2 was a .sam file that was used downstream to count mutations and compared to sequencing depth.

Bitvector generation, RNA and visualization

After mapping, a bitvector file was generated that counted the mutations. Each read was converted into a vector of '0' for matches, '1' for deletion, or 'A', 'T', C' or 'G' to indicate the identity of the mutation. Mates were compared, and a '?' or '.' replaced any ambiguous or missing bases. The bitvectors were then used to count mutations and normalized by sequencing depth and mutation rate in order to provide normalized DMS reactivity. We typically aimed for a minimum of 1000-fold coverage per base for population average DMS signal analysis for high-quality data.

The RNA structure prediction program RNAstructure was run using the RSample function in order to use the normalized DMS reactivities per base as constraints for folding[176]. This program produced both a visualization of the RNA secondary structure and bracket notation folding constraints. For final visualization, we used VARNA with the bracket notation constraints and color coded by DMS reactivity[177]. A downloadable version of the DMS-MaPseq pipeline can be found at rundmc.wi.mit.edu/cluster/dreem.

2.3 Results

HEK293t cells were transfected with HIV-1_{NHG}. The libraries were generated using 10 μg of total RNA, and rRNA was subtracted by using both the RNase H hybridization methods and RiboMinus. The final PCR products had an average size of 379 nt with a range of 253-496 for one representative library (Figure 2.3A), which corresponded with an average insert size of 259 nt. The libraries were sequenced and filtered for quality, then mapped to the HIV-1 genome in windows of 50-100 nt. The number of mutations per read for libraries was compared to the number of mutations per read for a site-specific PCR based approach for a region of the HIV-1 genome from the same starting total RNA.

The number of mutations per read from the PCR amplified read followed an approximately normal distribution around a mean of 4.42 mutations per alignment with a length of 243 nt (Figure 2.3B). The number of mutations per read for the library more closely approximated a Poisson distribution with an average of 0.55 mutations per alignment with a length of 100 nt. Since the DMS-modified RNA came from the same sample, part of the difference was likely due to the shorter read length in the library. When we aligned the PCR sample to a 100 nt window, we found that the average number of mutations per alignment was 1.98. We therefore speculate that the fragmentation step present in library generation enriches for RNA that has a low modification rate since modified RNA is susceptible to over-fragmentation.

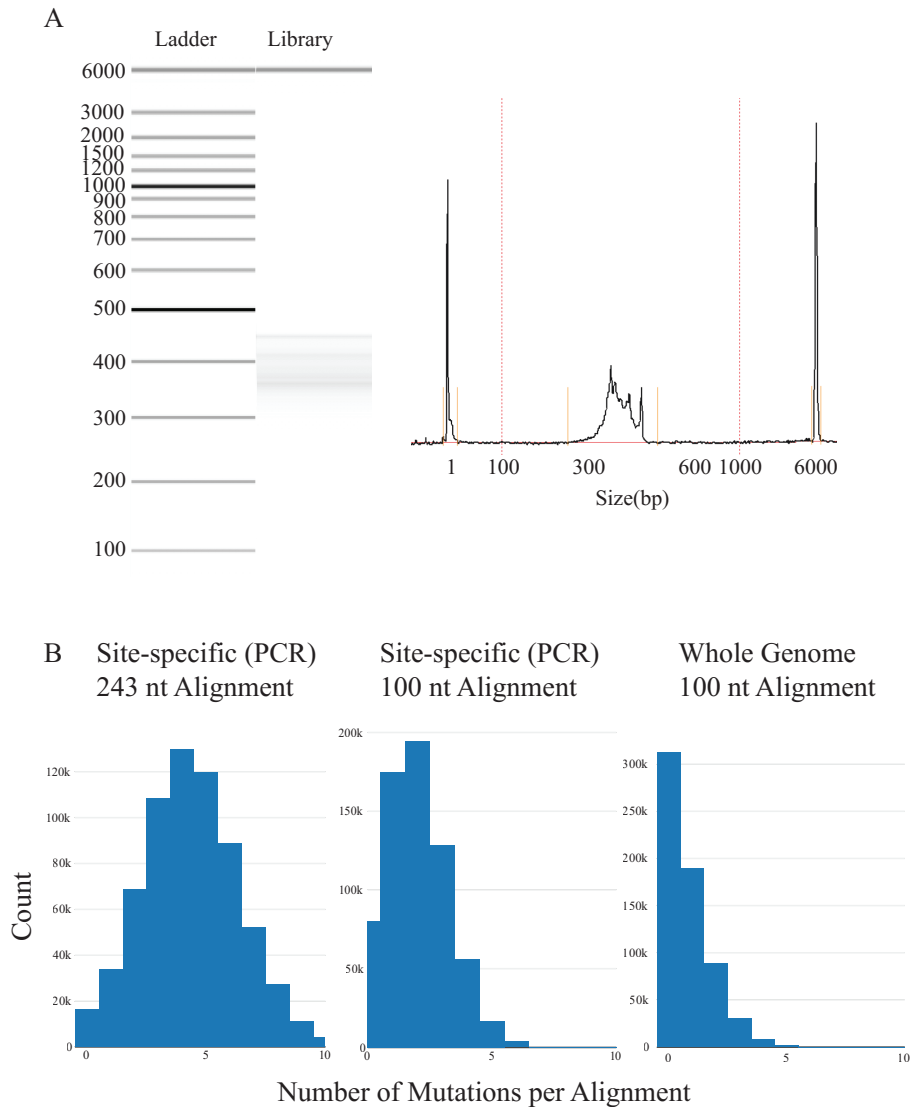

Figure 2.3*: Bioanalyzer trace of library generation and distribution of mutations per read. A) Bioanalyzer raw electrophoresis image and quantification for a fully constructed DMS-MaPseq library. B) On the left is a histogram of mutations per read for a site-specific PCR sample directed towards a region of the HIV-1 genome in HEK293t cells transfected with HIV-1$_{NHG}$ with an alignment length of 243 nt. In the center is a histogram of mutations per alignment for the same exact PCR sample aligned to a shorter window in the region of interest. On the right, the same starting DMS-modified RNA was used for library generation and the histogram of mutations per read is shown for the same region of the HIV-1 genome.*

The coverage of RNAseq reads aligning to the HIV-1 agreed with previous reports of HIV-1 RNAseq (Figure 2.4A)[178]. This result indicated that DMS treatment did not interfere with collection of the full range of HIV-1 RNA species. We observed a signal to noise ratio of ~10 across the HIV-1 genome (Figure 2.4B). This ratio was determined by comparing the average DMS mutation fraction of A's and C's in a 100 nt window to the mutation fraction of G's and U's. Only A and C should be modified by DMS. Some windows had a lower signal to noise ratio, but those windows were known to be highly structured, which would decrease the average DMS signal in that particular window.

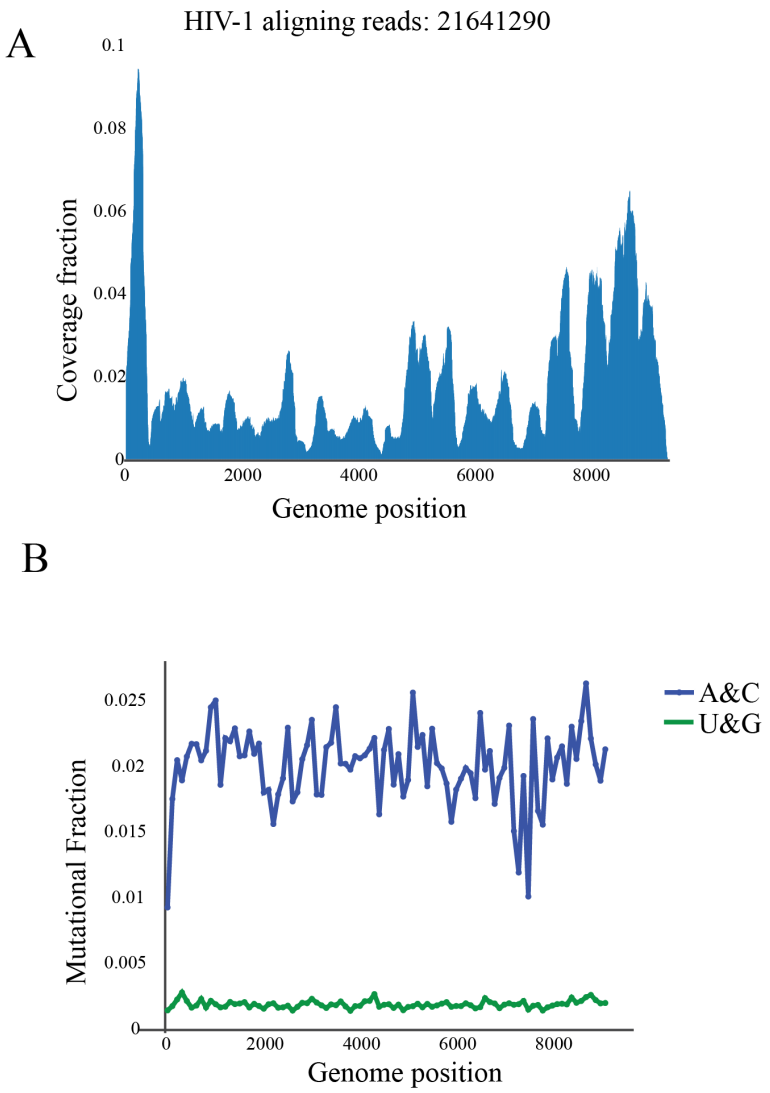

Figure 2.4*: Genome-wide HIV-1NHG library generation quality control. A) Coverage of HIV-1 genome with DMS-MaPseq data from HEK293t cells transfected with HIV-1NHG. B) Moving average of A and C mutational frequency in 100 nt windows after DMS-MaPseq compared to moving average U and G mutational frequency.*

Next, we viewed the first window of the HIV-1 genome, which contains the well-described TAR structure. We plotted the raw DMS-MaPseq data in a bar graph of mutational frequency per position of the genome. The only reactive bases were A's and C's, as is expected for quality DMS-MaPseq data (Figure 2.5A). The baseline mutational frequency (ie U and G) was ~10-fold

less than the mutational frequency for reactive bases. Next, we made a structural model with RNAstructure, using the DMS-MaPseq data on A and C bases as constraints (Figure 2.5B). The model was visualized with VARNA. The model of TAR recapitulated previously published structures[48,59,179] and represented an extremely stable RNA structure, with a free energy of -23.6 kcal/mol. Finally, we found that DMS-MaPseq data were highly reproducible. The Pearson R^2 value for two replicates of HEK293t cells transfected with HIV-1_{NHG} and DMS modified was 0.95 for the TAR region (Figure 2.5C).

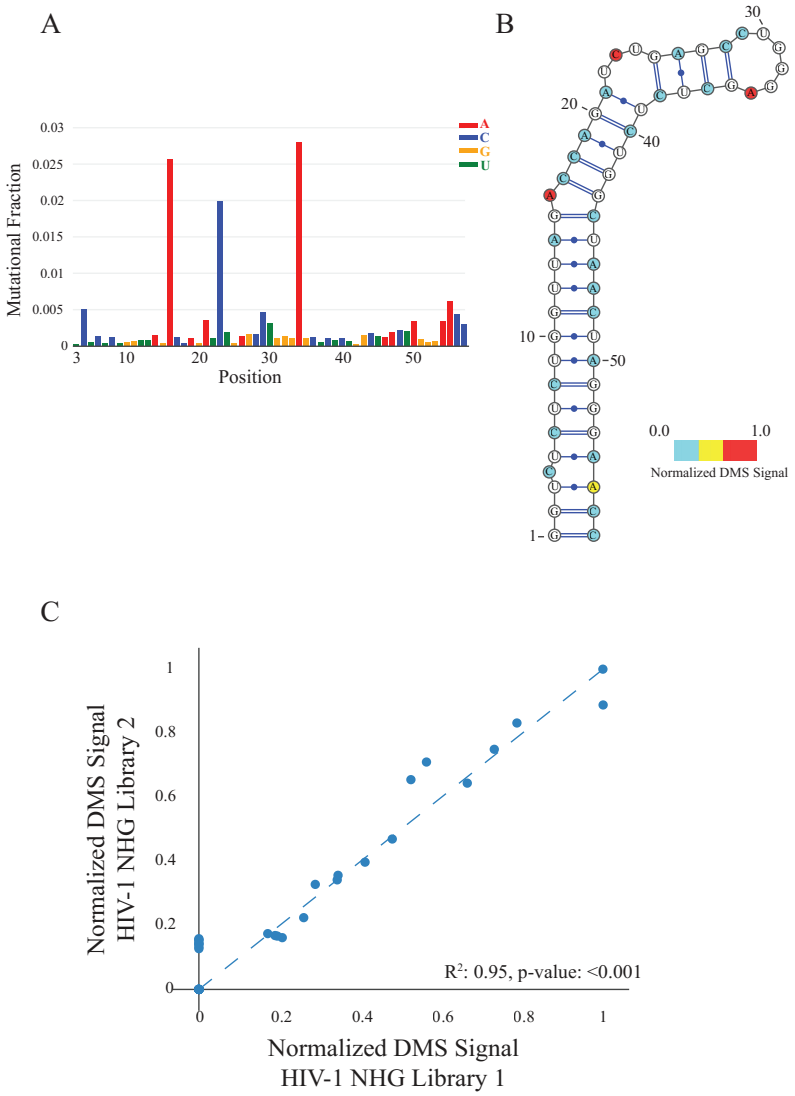

Figure 2.5*: DMS-MaPseq derived structural model for HIV-1 TAR. A) Raw data from a whole-genome library DMS-MaPseq for HIV-1 TAR from HEK293t cells transfected with HIV-1$_{NHG}$. The bar graph shows the mutational fraction, as a result of DMS-induced methylation, for all reads at each nucleotide position of the region of interest. Different nucleotides are color-coded. B) structural model for HIV-1 TAR based on the DMS-MaPseq constraints. The structural model was made using RNAstructure and visualized with VARNA. C) Scatterplot for replicates of TAR and immediate downstream sequence from two HEK293t and HIV-1$_{NHG}$ libraries.*

2.4 Discussion

RNA structure is capable of performing many biological functions, however it has been very difficult to study in cells. The flexibility of RNA makes most molecules intractable to study by typical biophysical techniques. Mutational profiling and sequencing protocols based either on SHAPE reagents or DMS have made an immediate impact on the ability to analyze RNA structure in the cellular environment, and in the context of the full-length RNA. One of the studies with the greatest impact to date has been study that combined SHAPE-MaP with psoralen cross-linking in order to identify intramolecular interactions thereby providing new insights into influenza virus segment packaging and reassortment. In that study, the authors were able to show unique hierarchies of SHAPE reactivities and intramolecular interactions for each strain of influenza. When strains were mixed within a cell, new reproducible interactions occurred between strains. The strength of the new interactions corresponded with the two strains' likelihood of reassortment[170].

Viral RNA structures may represent the most extreme examples of RNA structure usage, however functionally significant structures are likely ubiquitous in human RNAs as well. We hypothesize that human RNA structures are harder to identify because they only regulate functions in certain situations and are used in conjunction with many other factors. These features make the function of human RNA structures difficult to detect because any phenotypes will be weak. Viral RNA structures are used to conserve genomic space, and so are used regularly without many other factors. Therefore, phenotypic expression of RNA structures in viruses is easier to identify. For example, a human IRES was discovered shortly after the discovery of viral IRES[180]. This finding was initially viewed as a rare example, however 3-5% of

human genes can continue to be translated during extreme stress that prevents cap-dependent translation[181]. The list of human IRES elements continues to grow as more sophisticated and specific assays are developed to test the existence of each one. Because of the weaker structure, need for cofactors and situational usage within normal biology of human IRES's, they were not as easy to identify as viral IRES's. One reason it is important to study viral RNA structures is that they provide the blueprints to find more subtle human RNA structures.

Each viral system requires its own development of the DMS-MaPseq protocol. The host cells and structural components of the virus itself affect the ability of DMS to modify the RNA without causing excessive RNA degradation. In addition to finding the correct DMS concentration, RNA fragmentation is another critical troubleshooting step. The RNA is already fragmented during the DMS treatment, and over-fragmentation can lead to library generation failure or reads that are too short to be aligned properly. In this chapter, the technical issues for DMS-MaPseq of HIV-1 were solved. Although HIV-1 RNA structure is well-studied many questions remain that can now be addressed with DMS-MaPseq.

Chapter 3-Determination of RNA structural diversity and its role in HIV-1 RNA splicing regulation

Phillip J. Tomezsko[1,2,3]*, Vincent Corbin[4,5]*, Paromita Gupta[1]*, Harish Swaminathan[1], Margalit Glasgow[1,6], Sitara Persad[1,6], Matthew D. Edwards[7], Lachlan Mcintosh[4,8,9], Anthony T. Papenfuss[4,5,8,9,10], Ann Emery[11,12,13], Ronald Swanstrom[12,13,14], Trinity Zang[15], Tammy C.T. Lan[1], Paul Bieniasz[15,16], Daniel R. Kuritzkes[3,17], Athe Tsibris[3,17], Silvi Rouskin[1]

**These authors contributed equally to this work.*

[1]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, USA, [2]Program in Virology, Harvard Medical School, Boston, Massachusetts, USA, [3]Brigham and Women's Hospital, Boston, Massachusetts, USA, [4]Bioinformatics Division, Walter and Eliza Hall Institute, Parkville, Victoria, Australia,, [5]Department of Medical Biology, The University of Melbourne, Melbourne, Australia, [6]Massachusetts Institute of Technology, Cambridge, MA, USA, [7]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA [8]Peter MacCallum Cancer Centre, Melbourne, Australia, [9]Department of Mathematics and Statistics, University of Melbourne, Melbourne, Australia, [10]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, Australia, [11]Curriculum in Genetics and Molecular Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [12]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [13]Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [14]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [15]Laboratory of Retrovirology, The Rockefeller University, New York City, New

York, USA, [16]Howard Hughes Medical Institute, The Rockefeller University, New York City, New York, USA, [17]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA

V.C., H.S., M.G., S.P., M.E., L.M., T.P. and S.R. developed and wrote the DREEM clustering algorithm and analyzed validation studies. P.J.T. performed all cell and virus RNA modification assays. P.G. performed all *in vitro* RNA modification assays. P.J.T., P.G. and S.R. analyzed HIV-1 RRE and A3 RNA structure data. P.J.T., S.R., T.Z. and P.B. designed mutants. T.Z. produced mutant plasmids. A.E. and R.S. performed splicing analysis assays. P.J.T. generated the genome-wide DMS-MaPseq library. P.J.T., H.S., P.G. and S.R. analyzed genome-wide library data. T.C.T.L. conducted the U4/6 experiment. P.J.T. and S.R. wrote the manuscript. P.J.T., P.G., and S.R. created the figures. P.J.T., V.C., P.G., H.S., T.P., R.S., P.B., D.R.K., A.T. and S.R. edited the manuscript and figures.

This work was accepted by *Nature*, in press.

Abstract

HIV-1 must express all of its gene products from the same primary transcript, which undergoes alternative splicing to produce diverse protein products, including structural proteins and regulatory factors. Despite the critical role of alternative splicing, the mechanisms driving splice-site choice are poorly understood. Synonymous RNA mutations that lead to severe defects in splicing and viral replication indicate the presence of unknown cis-regulatory elements. DMS-MaPseq was used to probe the structure of HIV-1 RNA in cells. An algorithm called **D**etection of **R**NA folding **E**nsembles using **E**xpectation-**M**aximization (DREEM) was developed to reveal alternative conformations assumed by the same RNA sequence using the DMS-MaPseq data. Contrary to previous models which analyzed population averages, this study shows the widespread heterogeneous nature of HIV-1 RNA structure. In addition to confirming that *in vitro* characterized alternative structures for the HIV-1 Rev Responsive Element (RRE) exist in cells, this approach was used to discover alternative conformations at critical splice sites that influence the ratio of transcript isoforms. The simultaneous measurement of splicing and intracellular RNA structure provides evidence for the long-standing hypothesis that RNA conformation heterogeneity regulates viral splice site usage.

3.1 Introduction

Previous work on the genome-wide HIV-1 RNA structure *in vitro* and in virion provided a population average model, with the underlying assumption that every molecule within the population assumes the same conformation[59]. However, *in vitro* studies identified alternative conformations for the HIV-1 RRE and 5'UTR[45,52,53,56,182], raising the possibility that alternative structures have roles in viral RNA export from the nucleus and packaging in virions. Global synonymous mutations across the HIV-1 genome revealed cis-acting elements that impact splicing HIV-1 RNA[74]. Since the mutations were synonymous, it is possible that RNA structure contributes to splicing regulation. As splicing occurs on some HIV-1 RNA molecules but the majority remain unspliced[66], we hypothesized that alternative RNA structures were the unknown cis-elements. To resolve the fundamental question of whether RNA structure impacts splicing it is necessary to have the ability to distinguish multiple conformations for the same sequence in cells. We developed a clustering algorithm called **D**etection of **R**NA folding **E**nsembles using **E**xpectation-**M**aximization (DREEM) and demonstrated that we can quantitatively detect alternative structures.

DREEM starts with single molecule, chemical probing data, such as data from DMS-MaPseq. DMS adds methyl groups to unpaired adenine and cytosines of RNA molecules, which are converted to random mutations during cDNA synthesis with the RT enzyme TGIRT-III[150]. PCR amplifies the cDNA product and attaches sequencing adapters to the DNA, followed by massively parallel sequencing. Each resulting read is represented as a binary readout of mutations and matches, which is the input for DREEM. As DMS-MaPseq has negligible background error[141], the mutations observed on a single DNA molecule correspond to the DMS

accessible bases on the parent RNA molecule. The two key challenges for detecting heterogeneity are 1) DMS modification rates are relatively low (e.g. an open base has ~2-10% probability of being modified) and 2) the rate of DMS modification per open base is sensitive to the local chemical environment such that not all open bases are equally reactive to DMS. Traditional RNA structure determination approaches combine chemical probing data into population average signal per base, obscuring any underlying heterogeneity. In contrast, DREEM groups sequencing reads issued from each structure into distinct clusters by exploiting information contained in the observation of multiple modifications on single molecules. For instance, if two individual bases are never concurrently mutated on a single read, it follows that at least two conformations are present. DREEM identifies patterns of DMS-induced mutations on reads and clusters in a mathematically rigorous manner using an expectation-maximization (EM) algorithm (Figure 3.1**)**. The DMS modification rate per base for each cluster (or structure) is determined by iteratively maximizing a log-likelihood function to find and quantify the abundance of alternative structures directly from the dataset. The binary nature of the readouts allows for the use of a multivariate Bernoulli mixture model (MBMM) to compute the log-likelihood function[183]. The DMS modification pattern from each cluster is used to create a secondary structure model using RNAstructure[176]. We used DMS-MaPseq and DREEM to investigate alternative RNA structures across the HIV-1 genome and determine their role in HIV-1 splicing.

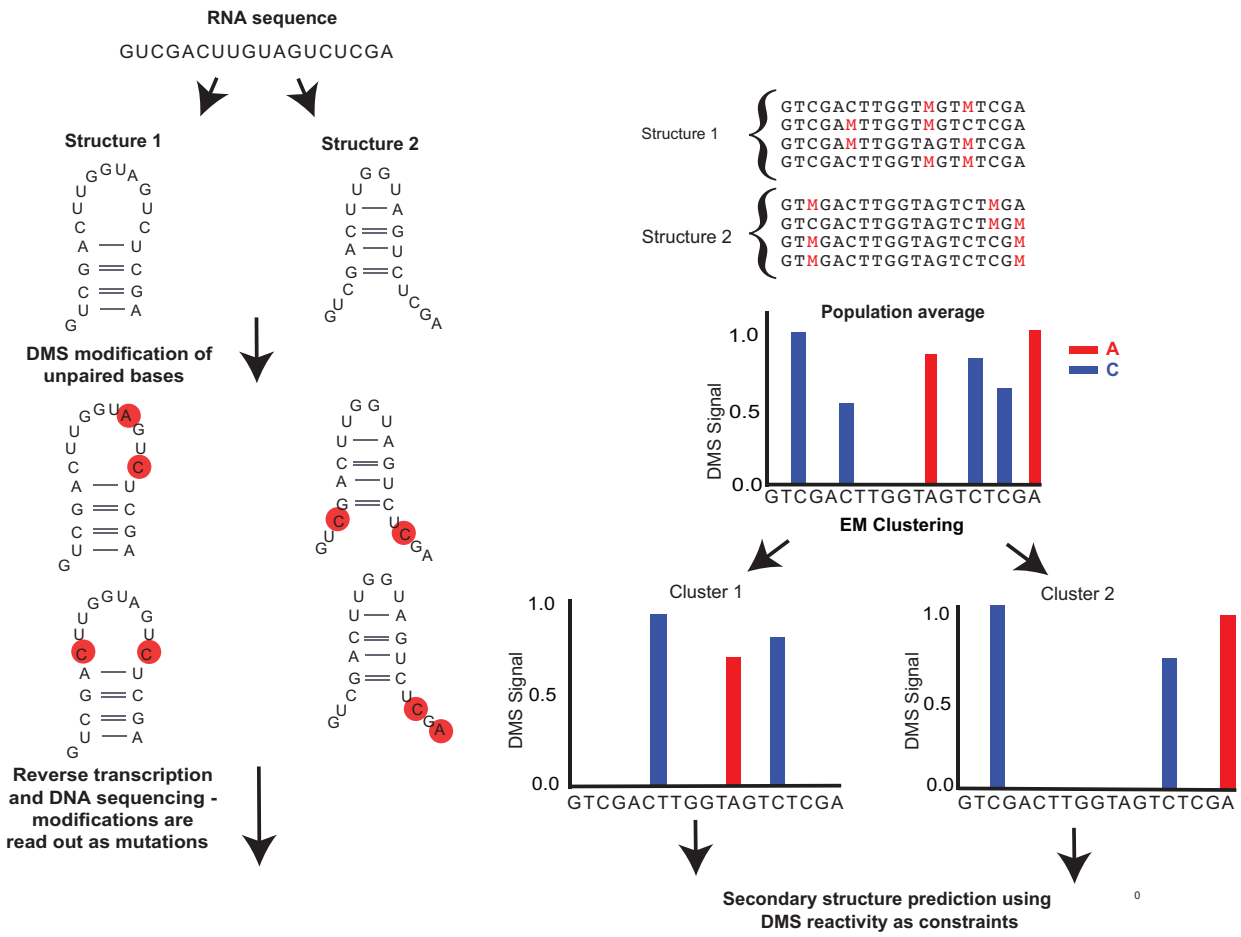

Figure 3.1*: Schematic of DMS-MaPseq and DREEM analysis of alternative RNA structure. A hypothetical RNA sequence that has two structures is chemically modified and sequenced. The DREEM algorithm clusters reads in order to make structural predictions of multiple conformations.*

3.2 Methods

DREEM clustering description

Relevant symbols and meanings:

- N: Total number of reads
- D: Length of region of interest in the reference
- $\boldsymbol{X} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}\}$: Set of all observed reads
- $\boldsymbol{S}$: Set of all allowed (observable) reads
- K: Number of clusters
- π_k: Mixing proportion of cluster k. $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ such that $\sum_{k=1}^{K} \pi_k = 1$.
- $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kD})$: Mutation profile of cluster k, where μ_{ki} is the mutation rate of base i in cluster k. $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$.
- y_{nk}: The latent Boolean variable representing the assignment of read n to cluster k.
- z_{nk}: The expectation of y_{nk}, or the probability that read n belongs to cluster k
- i, α: nucleotide index

The sequencing data from a sample were mapped to the corresponding reference genome using the Bowtie2 aligner[174]. The data observed $\boldsymbol{X}$ consisted of N reads $\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}\}$, each containing D nucleotides. Each read $\boldsymbol{x}_n \in \boldsymbol{X}$ represented a distinct RNA molecule that was DMS modified, reverse transcribed and amplified. The DMS modifications were read out as mutations. A read $\boldsymbol{x}_n$ could then be represented as a vector of D bits $(x_{n1}, \ldots, x_{nD})$, or a '*bit vector'*, where

$$x_{ni} = \begin{cases} 1, if\ base\ x_{ni}\ is\ mutated\ ; \\ 0, otherwise. \end{cases}$$

As DMS modification was not saturating (i.e. not every accessible base of a single molecule is modified), each open base in an RNA molecule had only a small probability (2-10%, depending on the DMS concentrations used) of being modified. As a consequence of this, a distinct mutation probability μ was associated with each base of the read. The mutation probabilities were assumed to be independent from each other. This assumption allowed each read to be considered as a random draw from a Bernoulli mixture model. In the event that the RNA molecules assume more than one structure, each structure appeared in the data as a collection of reads, or cluster, characterized by its own Bernoulli mixture model.

If K is the number of structures present in the sample, then the model is parameterized by:

a) The mutation probabilities $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$, where $\boldsymbol{\mu}_k = (\mu_{k1}, \dots, \mu_{kD})$ are the mutation probabilities of cluster k.
b) The mixing proportions $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_K\}$ of the K clusters, where π_k quantify the proportion of reads that belong to cluster k.

The EM algorithm used by DREEM for clustering assumes a Bernoulli mixture model[183]. Therefore, the probability of a base not being mutated in cluster k was: $\Pr(x_{ni} = 0|\boldsymbol{\mu}_k) = 1 - \mu_{ki}$, while the probability of a base being mutated in cluster k was: $\Pr(x_{ni} = 1|\boldsymbol{\mu}_k) = \mu_{ki}$. Hence the Bernoulli mixture model yielded the probability of observing a read $\boldsymbol{x}_n$ from cluster k as:

$$\Pr(\boldsymbol{x}_n|\boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}{}^{x_{ni}}(1 - \mu_{ki})^{1-x_{ni}} \quad (1)$$

Reads that contained mutations within three bases next to each other were very rare in the DMS-MaPseq dataset and occured at a frequency close to the sequencing error rate; i.e. the bit vectors 001001000, 001010000 and 001100000 were greatly underrepresented. This observation was likely due to the reverse transcriptase falling off the template when encountering adjacent methylations. Truncated reads did not get amplified during PCR and therefore were not represented when sequenced. To account for this bias, all rare reads containing mutations within three bases of each other were removed and the set of all reads with allowable mutations in $\{0,1\}^D$ that can be sequenced, $\boldsymbol{S}$, was computed. Therefore, equation (1) was modified as follows:

$$\Pr(\boldsymbol{x}_n|\boldsymbol{\mu}_k) = \frac{\prod_{i=1}^{D} {\mu_{ki}}^{x_{ni}}(1-\mu_{ki})^{1-x_{ni}}}{\sum_{\boldsymbol{x}'\in\boldsymbol{S}} \prod_{i=1}^{D} {\mu_{ki}}^{x_i'}(1-\mu_{ki})^{1-x_i'}}$$

In the initial step of the EM algorithm, the model parameters $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$ were randomly initialized. After the initialization of the parameters, the Expectation step and the Maximization step were executed one after the other in a loop until the log likelihood converges.

Two calculations were made in the Expectation step:

a) The responsibilities of the cluster were computed, i.e. the reads were assigned probabilistically to clusters:

$$z_{nk} = \frac{\Pr(\boldsymbol{x}_n|\boldsymbol{\mu}_k)\,\pi_k}{\sum_{j=1}^{K} \Pr(\boldsymbol{x}_n|\boldsymbol{\mu}_j)\,\pi_j}.$$

Here z_{nk} was the probability that read n belongs to cluster k. It could also be defined as the posterior probability, or responsibility, of cluster k given read n.

b) The expected complete-data log likelihood of observing the data X and latent variables $Y = \{y_{nk}\}$ given the model parameters was computed:

$$\mathbb{E}_{Y\sim Z} \ln \Pr(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \{\pi_k \Pr(\boldsymbol{x}_n|\boldsymbol{\mu}_k)\}$$

In the Maximization step, the model parameters were re-estimated by maximizing the expected value of the likelihood with respect to the parameters $\{\pi_k\}$ and $\{\mu_{ki}\}$.

a) Update mixing proportion of each cluster:

$$\pi_k = \frac{\sum_{n=1}^{N} z_{nk}}{N}$$

b) Update mutation profile μ_k of each cluster by solving the following system of equations for each k:

$$\frac{\sum_{x\in \boldsymbol{S}} x_\alpha \prod_{i=1}^{D} {\mu_{ki}}^{x_i}(1-\mu_{ki})^{1-x_i}}{\sum_{x\in \boldsymbol{S}} \prod_{i=1}^{D} {\mu_{ki}}^{x_i}(1-\mu_{ki})^{1-x_i}} = \frac{\sum_{n=1}^{N} z_{nk} x_{n\alpha}}{\sum_{n=1}^{N} z_{nk}} \quad \forall\, \alpha$$

These equations were derived by setting the derivatives of the expected complete-data log likelihood function to zero. After the EM clustering algorithm finished running, the reactivities of the bases in each cluster were given as inputs to RNAstructure for secondary structure prediction[176]. The DMS signal was normalized such that the median of the top ten most reactive positions is set to 1.0. To protect from spurious outliers, a 90% winsorization was used, effectively capping the reactivity at 1.0. Final visualizations of RNA secondary structure were created with VARNA[177].

Parameters used by the DREEM pipeline:

1. Minimum number of iterations of the EM algorithm to run before checking for convergence of the likelihood (*num_its*): 300
2. Number of EM algorithm runs (*num_runs*): 10. *num_runs* independent runs of the EM algorithm were carried out to ensure that the results from the algorithm are robust to the initialization of the model parameters and are repeatable.
3. Convergence threshold (*conv_thresh*): 1. The EM algorithm was stopped when $log(likelihood)_{iteration=n+1} - log(likelihood)_{iteration=n} < conv\_thresh$ after *num_its* iterations have been completed.
4. Signal threshold (*sig_thresh*): 0.005. Only mutation rates greater than *sig_thresh* were considered. All bases with a population average mutation rate less than *sig_thresh* were set to '0' in every bit vector.
5. $Bayesian\ Information\ Criterion\ (\text{BIC}) = \log(N) * D * K - 2\log(likelihood)$

   To test for over fitting the data we checked whether the EM algorithm passes two clusters by using the BIC test. If $BIC_{K=2} > BIC_{K=1}$, the algorithm stopped. Otherwise, the algorithm moved on to $K = 3$.
6. Bit vectors were filtered out if they do not satisfy one of the following four criteria:
   a. Informative bits threshold (*info_thresh*): 0.05-0.2. We set x_{ni} to '.' if base i is not covered by read x_n and to '?' if the base was of low quality (defined as having a Phred Quality Score less than 20). If the fraction of non-informative bits ('.', '?' and 'N') in the bit vector was greater than *info_thresh,* the bitvector was removed. After this filtering, all the non-informative bits were set to '0' in the remaining bit vectors.

b. Maximum number of mutations: If the number of mutations in the bit vector was greater than 3 times the standard deviation of the mutation distribution per read, the bit vector was removed
c. Invalid bit vectors: rare occurrences of bit vectors with adjacent mutations (within 3 nt) were considered to be part of background noise and were filtered out.
d. Rare instances where a bit vector consisted of a mutation ('1') right next to a non-informative base such as '.' and '?'.

7. Informative bases: Since DMS modifies only As and Cs, mutations at Ts and Gs were set to "0"s.

Cell Lines

HEK293t were obtained from ATCC. The cells tested negative for mycoplasma by LookOut Mycoplasma PCR Detection kit (Millipore-Sigma). The cells were maintained in Dulbecco's Modified Eagle Medium (ThermoFisher Scientific) supplemented with 10% heat-inactivated fetal bovine serum (FBS; ThermoFisher Scientific) and 100 U/mL penicillin/streptomycin (ThermoFisher Scientific).

Plasmid Construction

HIV-1 NL4-3 Infectious Molecular Clone (pNL4-3) was obtained from the NIH AIDS Reagent program. HIV-1_{NHG} is a full-length HIV-1 proviral plasmid, modified to replace a non-essential gene *nef* with *GFP* (Genbank accession code JQ585717.1). A Vpr-truncated derivative (Δvpr HIV-1_{NHG}) was constructed by generating an overlapping PCR with a C to T mutation and thus a

stop codon after Vpr amino acid 20. This PCR product was inserted into HIV-1_{NHG} using AgeI and SalI. All of the A3 splice site mutants were generated via overlapping PCR and inserted into a Δvpr HIV-1_{NHG}.

$CD4^+$ T Cell Isolation

Apheresis leukoreduction collars, obtained from the Brigham and Women's Hospital Crimson Core, were used to isolate peripheral blood mononuclear cell (PBMC) by density centrifugation using Lymphocyte Separation Medium (ThermoFisher Scientific). $CD4^+$ T cells were isolated by negative selection using EasySep Human $CD4^+$ T cell Enrichment Kit (StemCell Technologies). $CD4^+$ T lymphocytes were cultured at a density of approximately 1 million cells/mL in RPMI-1640 (ThermoFisher Scientific) medium supplemented with 10% fetal bovine serum (FBS) and 100 U/mL penicillin/streptomycin.

DMS Modification of In Vitro *Transcribed RNA*

gBlocks were obtained from IDT for the HIV-1 RRE, RRE MutA and MutB, control Structure 1, control Structure 2 and Adenosine deaminase (*add*) riboswitch. HIV-1 RRE and its mutants corresponded to nucleotides 7759-7990 based on HIV-1 vector pNL4-3 (Genbank accession code AF324493.1). *Add* corresponded to nucleotides 1590535-1590663 of *V.vulnificus* strain (Genbank Accession code CP037932.1). The U4/6 core-domain RNA construct was based on the interface of the U4 and U6 snRNA (Genbank accession code 2N7M_X). The gBlocks also contained 20-nt T7 RNA polymerase promoter sequence (TTCTAATACGACTCACTATA) on the 5′ end and a 23 nt sequence (CCGGAGTCGAGTAGACTCCAACA) on the 3′ end. The region of interest was amplified by PCR with a forward primer that contained the T7 promoter

sequence and a reverse primer complimentary to the 23 nt 3' sequence. The PCR product was used for T7 Megascript *in vitro* transcription (ThermoFisher Scientific) according to manufacturer's instructions. DNA template was degraded by adding 1 µL of Turbo DNase I (ThermoFisher Scientific) to the reaction and incubating at 37ºC for 15 minutes. The RNA was purified using RNA Clean and Concentrator -5 kit (Zymo). Approximately 1 µg of RNA was denatured at 95ºC for 1 minute. Based on the DMS concentration used in the next step, 300 mM sodium cacodylate buffer (Electron Microscopy Sciences) with 6 mM $MgCl_2$ was added so the final volume was 100 µl. The RNA was refolded by incubating for 20 mins at 37ºC. DMS (Millipore-Sigma) was added to achieve a final concentration of 0.25%-2.5% and incubated at 37ºC for 5 mins while shaking at 500 rpm on a thermomixer. The DMS was neutralized by adding 60 µL β-mercaptoethanol (Millipore-Sigma). The RNA was purified using RNA Clean and Concentrator -5 kit. For *in vitro* transcription of *add* riboswitch samples, one set of samples were incubated with 5 mM Adenine during the refolding stage at 37ºC.

$CD4^+$ T Cell Infection and DMS Modification

15 million $CD4^+$ T cells were activated by treatment with culture medium containing 10 µg/mL PHA (Millipore-Sigma) and 100 U/mL IL-2 (NIH AIDS Reagent Program; discontinued) for 72 hours. The cells were pelleted and infected with 200 µL of supernatant from HEK293t cells transfected with pNL4-3. After 48 hours, the supernatant was filtered with a 0.22 µM filter (Millipore-Sigma) and centrifuged at 28,000 x g for 1 hour, 4ºC in order to pellet virions. The cells were washed and resuspended in 15 mL of media and placed on a thermomixer at 37ºC. In order to modify the RNA, 200 µL of DMS, or ~1.3% v/v, (Millipore-Sigma) was added and the cells were incubated for 10 minutes while shaking at 800 RPM. DMS was neutralized by adding

30 mL of PBS (ThermoFisher Scientific) with 30% β-mercaptoethanol. The cells were centrifuged at 1000 x g for 5 mins, 4ºC. The cells were washed twice by resuspending the pellet with 15 mL of PBS with 30% β-mercaptoethanol and centrifugation to pellet. After washes, the pellet was resuspended in 1 mL of Trizol (ThermoFisher Scientific) and RNA was extracted following manufacturer's specifications. The virions were resuspended in 400 µL of PBS with 10 mM Tris pH 7 and 3 mM $MgCl_2$. 40 µL of DMS was added and the virions were incubated at 37ºC on a thermomixer while shaking at 800 RPM for 10 minutes. The DMS was neutralized with 400 µL of β-Mercaptoethanol and the RNA was purified using RNA Clean and Concentrator -5 kit. For unmodified RNA, 15 million CD4+ T cells were isolated and infected the same as described. 72 hours after infection, the supernatant was filtered with a 0.22 µM filter and virions were pelleted from the supernatant by centrifugation at 28,000xg for 1 hr, 4ºC and resuspended in 1 mL of Trizol. The cells were pelleted, resuspended in 1 mL of Trizol and RNA was extracted following manufacturer's instructions.

HEK293t Transfection and DMS Modification

HEK293t cells per well were seeded on a 6-well plate at a concentration of 0.9 million cells/well and incubated overnight. The cells were transfected using 2 µg of plasmid DNA (pNL4-3, pNHG or mutant) per well with X-tremeGENE 9 (Millipore-Sigma) following manufacturer's [instructions and incubated for 48 hours. After incubation, virions were collected from the supernatant and DMS modified as above. The cells were washed with PBS and 2 mL of culture medium with ~1.3% v/v DMS was added to each well. The plates were incubated at 37ºC for 4 mins. The medium containing DMS was immediately removed and replaced with 2 mL/well of PBS with 30% β-mercaptoethanol. Cells were scraped and centrifuged at 1000 x g for 5 mins,

4ºC. The pellet was resuspended in PBS and centrifuged to pellet twice. The pellet was resuspended in 1 mL of Trizol and RNA was extracted following manufacturer's specifications. For unmodified RNA, HEK293t were plated and transfected with the same protocol as above. 48 hours after transfection, the supernatant was filtered with a 0.22 μM filter and virions were pelleted from the supernatant by centrifugation at 28,000xg for 1 hr, 4ºC and resuspended in 1 mL of Trizol. The cells were trypsinized, washed and resuspended in 1 mL of Trizol. RNA was extracted following manufacturer's instructions.

RT-PCR with DMS-modified RNA from Cells or In Vitro *Transcription*

rRNA was subtracted from 1-3 ug of RNA per reaction. In order to subtract rRNA, 1 μL of rRNA subtraction mix (3 μg/ul) and 2.5 μL of 5x hybridization buffer (1 M NaCl, 500 mM Tris-HCl pH 7.5) were added to each reaction, and the final volume adjusted with nuclease-free water to 12.5 μL. The samples were incubated at 68ºC and the temperature was reduced by 1ºC/min until the reaction was at 45ºC. To degrade the RNA:DNA duplexes, 5 μL of RNase H buffer and 2 μL of Hybridase thermostable RNase H (Lucigen) were added and nuclease-free water was added until the final volume was 40 μL. The samples were incubated at 45ºC for 30 mins. The RNA was cleaned with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of fragments >200 nt and eluted in 45 μL of nuclease-free water. DNA was degraded by adding 5 μL of Turbo DNase buffer and 1 μL of Turbo DNase (ThermoFisher Scientific) to each reaction and incubated for 30 mins at 37ºC. To stop DNA degradation, 5.1 μL of DNase inactivation reagent (ThermoFisher Scientific) was added and incubated 5 mins at room temp with intermittent manual mixing. The RNA was cleaned with RNA Clean and Concentrator -5 following instructions for recovery of fragments >200 nt and eluted in 15 μL of

nuclease-free water. For reverse transcription, 1 µL of RNA was added to 3.5 µL of nuclease-free water, 2 µL of 5x First Strand buffer (ThermoFisher Scientific), 1 µL of 10 µM reverse primer, 1 µL of dNTP, 0.5 µL of 0.1M DTT, 0.5 µL of RNaseOUT, and 0.5 µL of TGIRT-III (Ingex). The RT reaction was incubated at 57°C for 1.5 hours, followed by a 5 mins at 80°C. To degrade the RNA, 1 µL of RNase H (New England Biolabs) was added to the RT reaction and incubated for 20 mins at 37°C. PCR was performed to amplify the samples using either Advantage HF 2 DNA polymerase (Takara) or Phusion (NEB) for 25-30 cycles according to manufacturer's specifications. PCR product was purified by QIAquick PCR purification (Qiagen) and sequenced either on MISeq or iSeq100 (Illumina) to produce either 100 nt single-end reads or 150x150 nt paired-end reads respectively.

Library Generation with DMS-modified RNA for HIV-1 genome RNA structure

A total of 10 µg extracted DMS-modified RNA from HEK293t transfected with NHG plasmid was split into 3 reactions for the first step of RNase H-based rRNA subtraction. The steps for RNase H and DNase treatment mentioned above were followed. After DNase treatment, the three reactions were eluted in 8.5 µL of nuclease-free water and combined. An additional rRNA subtraction step was performed using the RiboZero Human/Mouse/Rat rRNA removal kit (Illumina; discontinued) according to manufacturer's specifications. After RiboZero, the RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of fragments >200nt and eluted in 10 µL of nuclease-free water. The RNA was fragmented using the RNA Fragmentation kit (ThermoFisher Scientific) with a fragmentation step of 45 seconds at 70°C. The RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of all fragments and eluted in 6.5 of µL

nuclease-free water. In order to dephosphorylate the RNA ends, 1 µL of CutSmart buffer (New England Biolabs), 1.5 µL of Shrimp Alkaline Phosphatase (New England Biolabs) and 1 µL of RNaseOUT (ThermoFisher Scientific) were added and incubated at 37ºC for 1 hour. Linker was ligated to the RNA by adding 6 µL of 50% PEG-800 (New England Biolabs), 2.2 µL of 10x T4 RNA Ligase buffer (New England Biolabs), 2 µL of T4 RNA Ligase, truncated KQ (England Biolabs) and 1 µL of 10 µM linker and incubated for 18 hours at 22ºC. The RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of all fragments and eluted in 15 µL of nuclease-free water. Excess linker was degraded by adding 2 µL of 10x RecJ buffer (Lucigen), 1 µL of RecJ exonuclease (Lucigen), 1 µL of 5'Deadenylase (New England Biolabs) and 1 µL of RNaseOUT, then incubating for 1 hour at 30ºC. The RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery fragments > 200 nt and eluted in 11 µL of nuclease-free water. For reverse transcription, 1 µL of RT primer, 1 µL of 0.1 M DTT, 4 µL of 5x First Strand buffer, 1 µL of dNTP, 1 µL of RNaseOUT and 1 µL of T-GIRT III were added and the sample was incubated for 2 hours at 65ºC. RNA was degraded by adding 1 µL of 4 N NaOH and incubating at 95ºC for 3 mins. The RT product was mixed with an equal volume of 2x Novex TBE-Urea sample buffer (ThermoFisher Scientific) and run on a 10% TBE-Urea gel (ThermoFisher Scientific). The ~300-400 nt cDNA was extracted. The purified cDNA was circularized using the CircLigase ssDNA Ligase kit (Lucigen). For PCR amplification, 2 uL of the circularized product was used for PCR using Phusion. The sample was run for a maximum of 14 cycles. Following PCR, the product was run on an 8% TBE gel and the ~350-450 nt DNA product was gel extracted. The final PCR product was quantified by Bioanalyzer (Agilent). The product was then sequenced by Novaseq S4 (Illumina) to produce 150x150 nt paired-end reads. The same library generation protocol was

followed for *in vitro* transcribed and DMS-modified U4/6 core-domain with some modifications. The starting amount of U4/6 core-domain RNA was 250 ng of RNA as part of a pool of RNA totaling 4 μg. No fragmentation or rRNA removal were performed for U4/6 core-domain library generation.

HIV-1 Splice Junction Usage Analysis

Splice analysis was performed according to a previously written protocol[64]. Briefly, two separate RT-PCR reactions were performed with 2 μg of total unmodified RNA from HEK293t cells transfected with plasmid containing HIV-1_{NHG}, Δvpr HIV-1_{NHG}, and HIV-1 mutants. One reaction was designed to reverse transcribe all HIV-1 multiply spliced products with a reverse primer that spans the D4A7 splice junction. The second reaction was designed to reverse transcribe HIV-1 singly spliced mRNA with a reverse primer lies in the *env* intron. The forward primer used in both PCR reactions is located upstream of D1. Reverse transcription was performed with SuperScript III (Thermo Fisher Scientific) at 55°C for 1 hour followed by 15 minutes at 70°C. RNA was degraded by adding 1 μL of RNase H and incubating at 37°C for 20 minutes. The cDNAs were then purified with Agencourt RNACleanX beads at a ratio of 2:1 (Beckman Coulter). Two successive rounds of PCR were used to add adapters for sequencing using the KAPA robust PCR kit (KAPA Biosystems). The first PCR uses with a forward primer that is located in the shared upstream D1 sequence that also has an adapter. The second round adds the universal adapter and Illumina indexed sequencing primers. The PCR products were then sequenced by Illumina Miseq, 300x300 nt paired-end reads.

Statistical Methods

Statistical analysis of DREEM clusters was quantified by Pearson's correlation.

Library Linker and Primers

All oligos were ordered from IDT.

StemA/StemC T7 forward primer: TAATACGACTCACTATAGAAAGGATCGG

StemA/StemC T7 reverse primer: ATCCCAGCGCGTGGTGCA

StemA/StemC RT primer: ATCCCAGCGCGTGGTGCA

StemA/StemC PCR forward primer: GAAAGGATCGGAAGACTCCACAG

StemA/StemC PCR reverse primer: ATCCCAGCGCGTGGTGCA

Add riboswitch T7 forward primer:

TTCTAATACGACTCACTATAGGACACGACTCGAGTAGAGTCG

Add riboswitch forward primer: GACACGACTCGAGTAGAGTCG

Add riboswitch reverse primer: TGTTGGAGTCTACTCGACTCCGGT

HIV-1 RRE T7 forward primer: TAATACGACTCACTATAGGAGCTTTGTTCC

HIV-1 RRE T7 reverse primer: GGAGCTGTTGATCCTTTAGGTATCTTTC

HIV-1 RRE RT primer: GGAGCTGTTGATCCTTTAGGTATCTTTC

HIV-1 RRE PCR forward primer: GGAGCTTTGTTCCTTGGGTTCTTGG

HIV-1 RRE PCR reverse primer: GGAGCTGTTGATCCTTTAGGTATCTTTC

HIV-1 A3 PCR forward primer: TGAAACTTACGGGGATACTTGGGCAGGA

HIV-1_{NL4-3} A3 PCR and RT reverse primer:

GAAGCTTGATGAGTCTGACTGTTCTGATGAGC

HIV-1_{NHG} A3 PCR and RT reverse primer: CTTCGTCGCTGTCTCCGCTTCTTCC

To generate Δvpr HIV-1_{NHG}:

NL AgeF: AGC TAG AAC TGG CAG AAA ACA GGG AGA TTC

NL SalIR: CCA TTT CTT GCT CTC CTC TGT CGA GTA ACG C

dVprS: GGA AAC TGA CAG AGG ACA GAT GGA ATA AGC CCC AGA AGA CC

dVpr AS: GGT CTT CTG GGG CTT ATT CCA TCT GTC CTC TGT CAG TTT CC

To generate A3 Splice Site Mutants:

NL 5599F: CATACAATGAATGGACACTAGAGCTTTTAG

NL BamHIR: CGTCCCAGATAAGTGCCAAGGATCCGTT

A3SLMut1 S:

TCCATTTCAGAATTGGGTGTCGAGTAAGCCTAATAGGCGTTACTCGACAGAGGA

A3SLMut1 AS:

TCCTCTGTCGAGTAACGCCTATTAGGCTTACTCGACACCCAATTCTGAAATGGA

A3SLMut 2 S: GAATTGGGTGTCGACAACGCCTAATAGGCGTTACTCGAC

A3SLMut2 AS: GTCGAGTAACGCCTATTAGGCGTTGTCGACACCCAATTC

A3SLMut3 S: GGTGTCGACATAGCAGAATCTGCTATACTCGACAGAGGAGAGCAA

A3SLMut3 AS: GGTGTCGACATAGCAGAATCTGCTATACTCGACAGAGGAGAGCAA

A3SLMut4 S: TCAGAATTGGGTGTCGAAACAGCGAAATAGGCGTTACTCGACAGA

A3SLMut4 AS: TCTGTCGAGTAACGCCTATTTCGCTGTTTCGACACCCAATTCTGA

A3SLMut5 S:

TCAGAATTGGGTGTCGAAACAGCGAAATTCGCGTGTTTCGACAGAGGAGAGCAA

A3SLMut5 AS:

TTGCTCTCCTCTGTCGAAACACGCGAATTTCGCTGTTTCGACACCCAATTCTGA

Library generation linker:

/5rApp/TCNNNNNNNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGA/3ddC/

Library generation RT primer:

/5Phos/AGATCGGAAGAGCACACGTCTGAACTCCAG/iSp18/TCTTTCCCTACACGACGC

TCTTCCGATCT

Library generation forward PCR primer:

CAAGCAGAAGACGGCATACGAGAT**XXXXXX**GTGACTGGAGTTCAGACGTGTGCTC

Library generation reverse PCR primer:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC

Splice Analysis, Multiply Spliced Reverse Primer:

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNNNNNNNCAGTTCG

GGATTGGGAGGTGGGTTGC

Splice Analysis, Singly Spliced Reverse Primer:

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNNNNNNNGTACTATA

GGTTGCATTACATGTACTACTTAC

Splice Analysis, PCR Round 1 Forward Primer:

GCCTCCCTCGCGCCATCAGAGATGTGTATAAGAGACAGNNNNTGCTGAAGCGCGC

ACGGCAAG

Splice Analysis, PCR Round 2 Reverse Primer:

CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGCTC

Splice Analysis, PCR Round 2 Forward Primer:

AATGATACGGCGACCACCGAGATCTACACGCCTCCCTCGCGCCATCAGAGATGTG

Data Availability

Sequencing data can be obtained from the GEO database using accession number GSE131506.

Software and Code Availability

Sequence alignment: Bowtie2 2.3.4.1. For code development: python v. 3.6.7. For read trimming: TrimGalore 0.4.1. For read quality assessment: FastQC v0.11.8. For RNA secondary structure analysis: RNAstructure v6.0.1. For calculating post-mapping statistics: Picard 2.18.7. RNA secondary structure visualization: VARNA v3.93. HIV-1 splicing analysis: https://github.com/SwanstromLab/SPLICING. Splice plot creation: R version 3.5.1. For figure construction: Adobe Illustrator CC 2019. For data analysis: Microsoft Excel 2018. Plot generation: Plotly v3.2.1. DREEM clustering algorithm is available at https://gitlab.com/Corbin-Rouskin/RNA_structure/tree/gpu.

3.3 Results

Development and validation of DREEM algorithm

Control experiments on denatured RNA indicated that TGIRT-III is unable to read-through mismatches located within 3 nt of each other. In order to account for this observation, the standard MBMM log-likelihood function was modified. Upon convergence of the clustering, the DMS-signal from each cluster was used as a constraint in RNAStructure[176]. DREEM is unique among algorithms for RNA folding ensembles[184] because DREEM directly clusters the experimental data in contrast to clustering on structural predictions of population average data. Clustering before secondary structure model generation allows for the discovery of novel RNA structures, in contrast to previous work on the RING-MaP technique[185,186]. Purely computational algorithms rely on suboptimal folding to create variation not captured by minimum free energy calculations. However, using experimentally derived constraints is superior to using randomly generated constraints[187,188]. Moreover, DREEM does not rely on thermodynamics for detecting and identifying alternative conformations, and therefore can be used on *in vivo* data to model RNA folding in the presence of cellular factors.

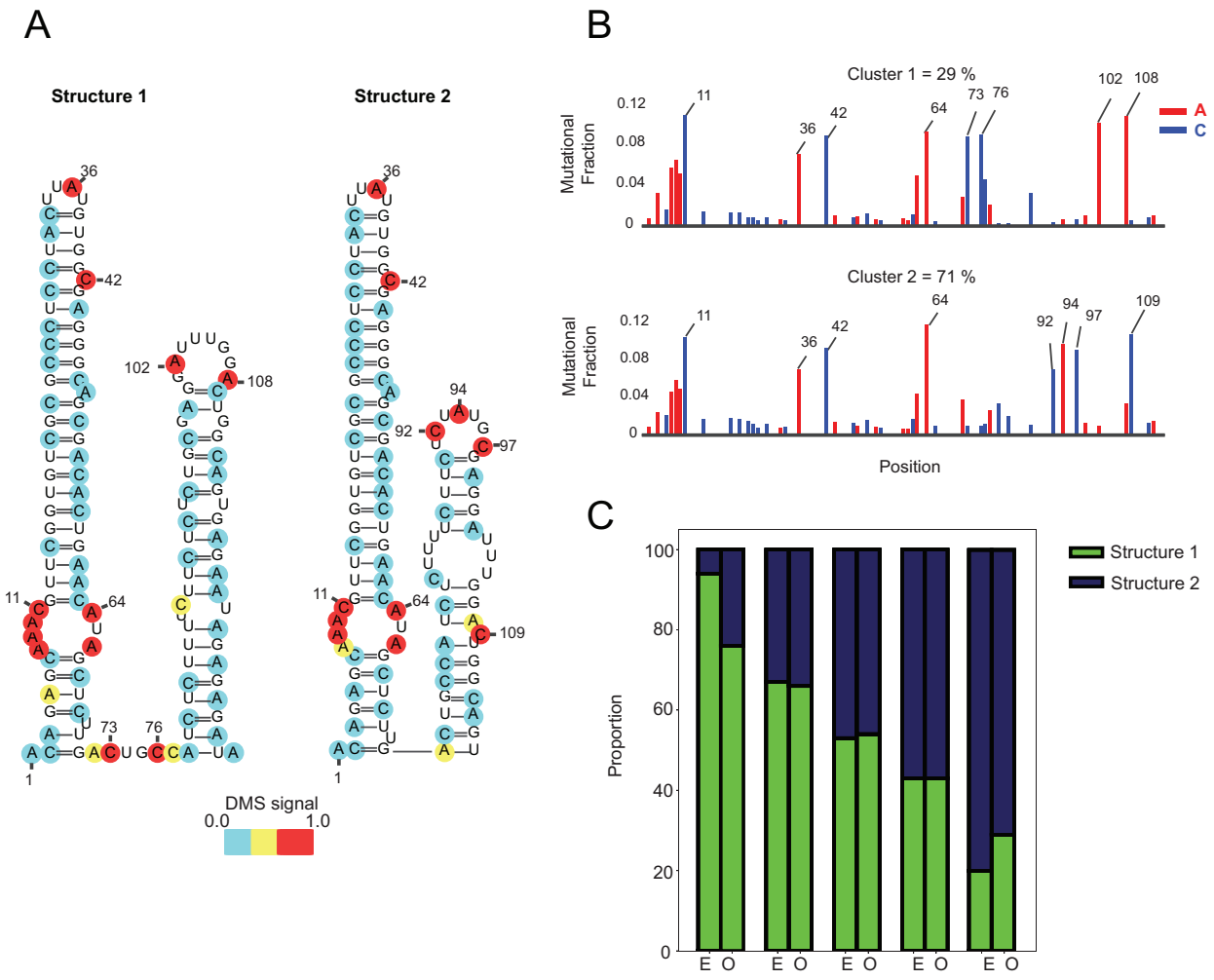

Figure 3.2*: DREEM algorithm validation using mixing of two known structures. A) Structural prediction from DMS-MaPseq probing of* in vitro *transcribed and modified Structure 1 and Structure 2. B) Mutational fraction for each nucleotide after DREEM clustering for up to two clusters. C) Expected (E) vs observed (O) cluster detection for mixed Structure 1 and Structure 2 after DMS-MaPseq and DREEM analysis.*

To validate DREEM, two RNA molecules were *in vitro* transcribed and DMS-modified that are nearly identical in sequence but form different structures (Structure 1 and Structure 2). These sequences were designed based on the RiboSNitch in the human gene *MRPS21*, which forms alternative structures based on a single nucleotide polymorphism[189]. Mixtures of the two RNAs were experimentally produced in varying proportions and used to generate DMS-MaPseq data.

DREEM clustered the DMS-MaPseq data and successfully identified the two structures down to a mixing ratio of 94%:6% (Figure 3.2). We also tested DREEM using *in vitro* transcribed, folded and DMS-modified adenosine deaminase (*add)* riboswitch from the bacterium *Vibrio vulnificus*, which undergoes a conformational shift upon binding of adenine[190-192]. We found that *add* structures that promote translation, which are stabilized by adenine, went from 18% to 89% upon addition of 5 mM adenine (Figure 3.3A). Structural predictions based on DREEM output for *add* in the absence of 5 mM adenine matched previously identified structures (Figure 3.3B)[190-192].

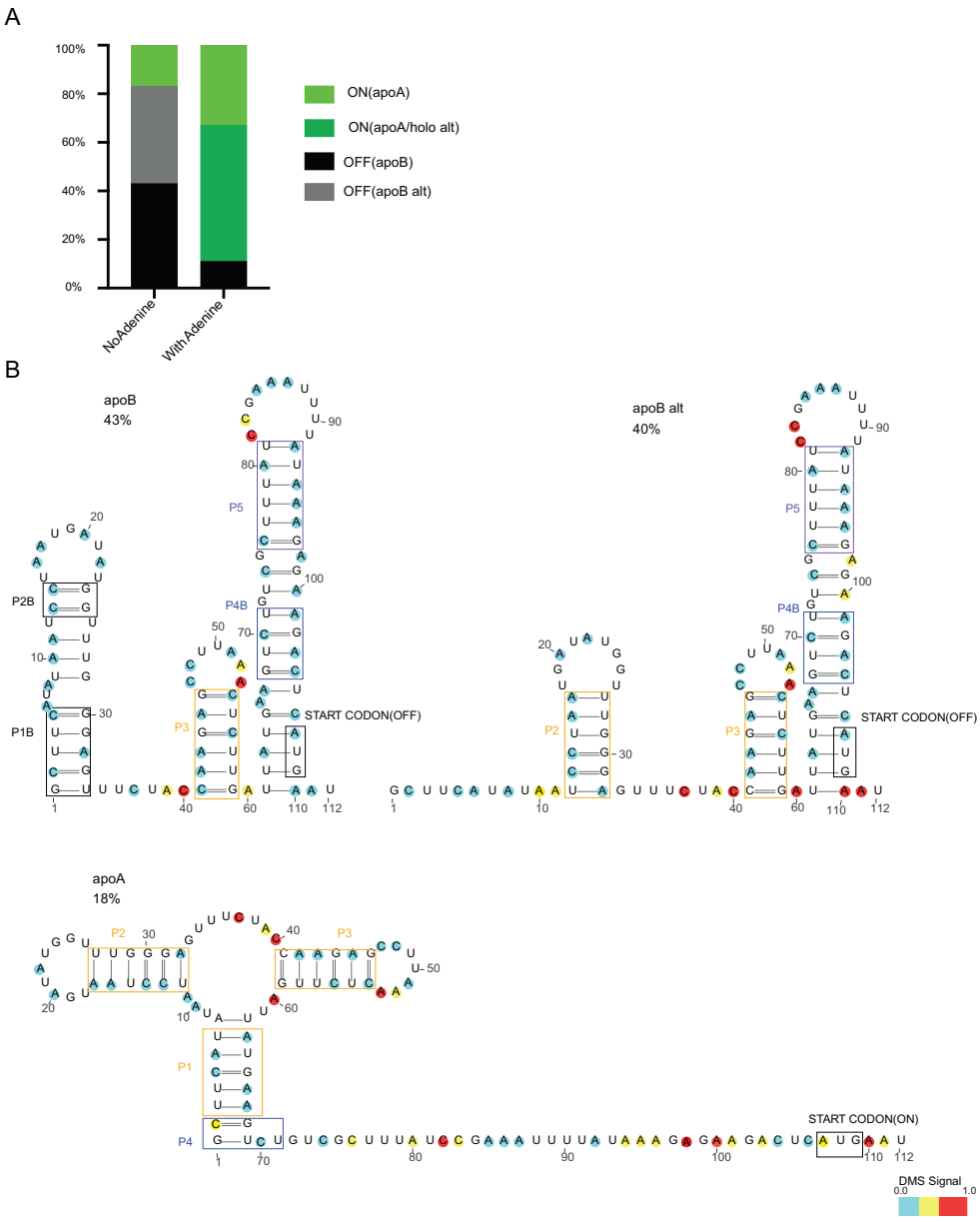

***Figure 3.3**: Adenine riboswitch (add) alternative RNA structure in presence or absence of adenine. A) The ratios of alternative structures detected by DREEM of* add *while being folded in the presence or absence 5 mM adenine. The 'on' or 'off' form are based on observations based on previous studies. B) Structural predictions of alternative RNA structures detected by DREEM of* add *in the absence of adenine.*

Intracellular alternative HIV-1 RRE structure is consistent with in vitro *and in virion structures*

The RRE of HIV-1 is multi-stem RNA structure that binds to the viral protein Rev and allows for the nuclear export of unspliced and partially spliced HIV-1 RNA. Previous studies physically separated distinct RNA conformations by native gel electrophoreses and revealed two alternative structures for RRE *in vitro*: a 5-stem and 4-stem structure. Specific mutations are able to stabilize either of the alternative conformations[52]. DREEM accurately identified the DMS signal for mixtures of *in vitro* transcribed, folded and DMS-modified RNA from the 5-stem (MutA) and 4-stem (MutB) mutant structures and robustly quantified their mixing ratios (Figure 3.4A). *In vitro* transcribed, folded and DMS-modified wildtype HIV-1_{NL4-3} RRE sequence was found to be in a mixture of ~27% 4-stem and ~73% 5-stem structure after DREEM analysis.

DMS-MaPseq/DREEM was applied to the study of HIV-1 RRE structure in primary cells, which is possible as DMS is cell membrane permeable[193]. Activated $CD4^+$ T cells were infected with HIV-$1_{NL4-3.}$ We performed chemical probing *in vivo* and in virions. The RRE sequence forms the same alternative structures regardless of the environment (*in vitro*, *in vivo*, and in virion), favoring the 5-stem fold (Figure 3.4B). Both the 5-stem and 4-stem were detected in each condition. The percentage of 5-stem ranged from 65%-73% and the percentage of 4-stem was 27%-35%.

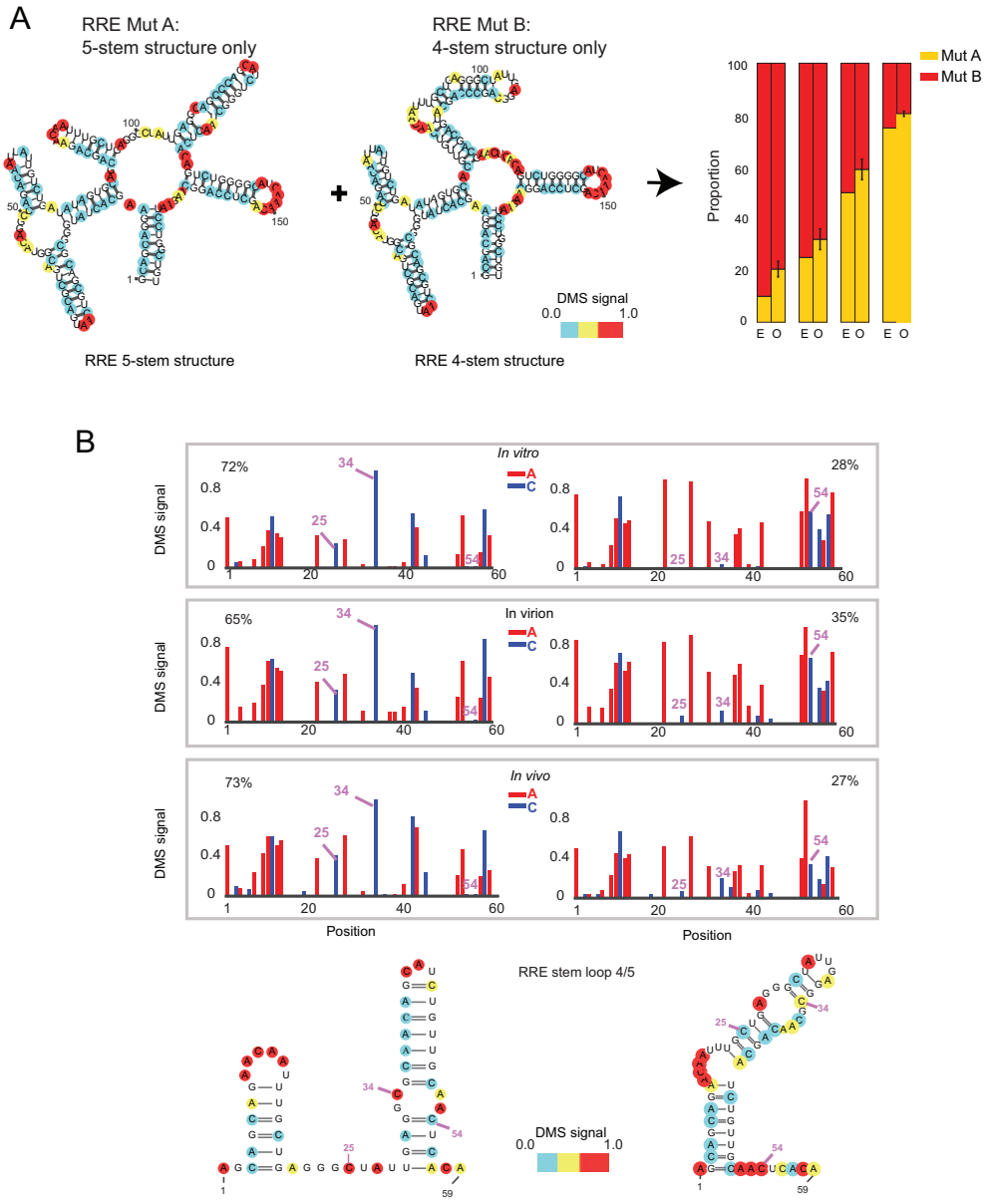

Figure 3.4*: Alternative HIV-1 RRE structures detected in different folding environments. A) MutA and MutB structure predictions based on DMS-MaPseq data from* in vitro *transcribed and modified RNA. The bar graph shows the expected (E) versus observed (O) proportions of 5-stem versus 4-stem structure after mixing. B) Normalized DMS signal for the Stem III/IV region of HIV-1$_{NL4-3}$ RRE modified* in vitro*, in virions and in cells. DREEM structural predictions for Stem III/IV region shown from in cell modified RNA.*

Alternative RNA structure at the HIV-1 A3 splice acceptor site influences splice usage

We next examined the role of RNA structure in HIV-1 splicing. Alternative splicing is the major mechanism used by HIV-1 to express all of its gene products from a single type of pre-mRNA (i.e. full-length genomic viral RNA). Splice site usage must be regulated to produce the correct proportion of HIV-1 transcripts. HIV-1 transcripts spliced at the A3 acceptor splice site are the only source of mRNAs for the viral transcriptional activator Tat[31].

DREEM detected alternative RNA structures at the HIV-1_{NL4-3} A3 splice acceptor site. The structures that formed at the A3 splice site in $CD4^+$ T cells differ drastically from previously proposed models based on population average data[59]. Strikingly, the two main conformations identified by DREEM either occlude (~40%, Cluster 1) or expose (~60%, Cluster 2) the polypyrimidine tract where U2AF heterodimer binds and A3 splice site (abbreviated together as A3ss, Figure 3.5A). We termed the occluded structure A3 stem-loop (A3SL). The A3SL is not specific to the HIV-1_{NL4-3} and forms in HIV-1_{NHG} in HEK293t cells (Cluster 1, Figure 3.5B). Notably, strong heterogeneity at the A3ss folded *in vitro* was detected, and the A3SL could be identified after clustering as in the intracellularly modified RNA. Although the A3SL cluster matched between the two conditions, the other cluster did not have the same structure.

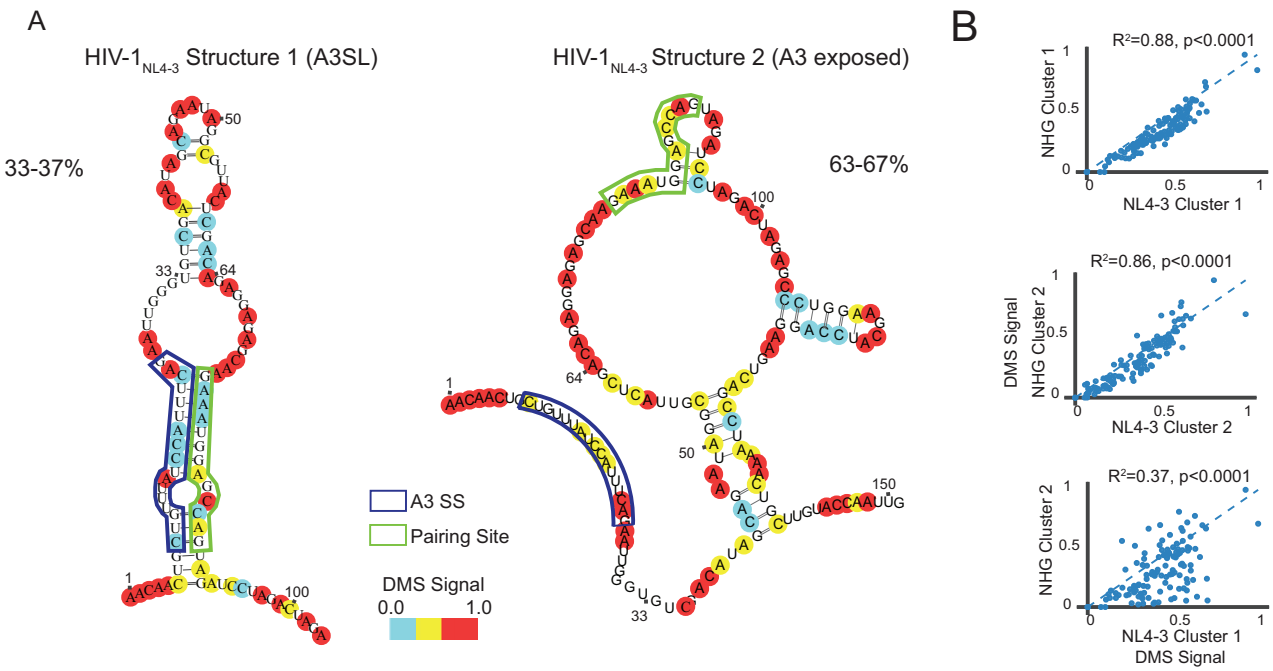

Figure 3.5*: HIV-1$_{NL4-3}$ A3 splice acceptor site alternative RNA structures detected by DREEM. A) DMS-modified RNA from CD4$^+$ T cells infected with HIV-1$_{NL4-3}$ was clustered by DREEM. Two clusters were identified and are shown with the location of the A3ss and proportions of clusters. B) Scatterplot comparing clusters of CD4+ T cells infected with HIV-1$_{NL4-3}$ and HEK293t cells transfected with HIV-1$_{NHG}$.*

To perturb the population of RNA structures and measure the effect on splicing, we took advantage of A3ss location in the *vpr* coding region, which is dispensable for growth in cell-culture. A strain with a pre-mature stop codon in *vpr* was utilized to ensure that observed effects were not due to loss of function of *vpr* (Δvpr HIV-1$_{NHG}$). To test the effect of structure on splicing, mutations distal from the splice site sequence were designed. These mutants avoided known protein-binding regions, including splice enhancer and splice inhibitory elements. Mutants A3SLMut1, 2, and 3 were predicted to thermodynamically stabilize A3SL and decrease splicing at A3ss (Figure 3.6A). Using a deep-sequencing based HIV-1 splicing assay[64], all three stabilizing mutants were found to have a lower rate usage of A3ss (Figure 3.6B), significantly decreasing expression of *tat* transcripts relative to background strain. The stabilizing prediction

was experimentally determined for A3SLMut1 and it was found that indeed the A3SL structure increased in abundance. In addition, the introduced mutations allowed for the formation of a new alternative conformation in ~35% of molecules, underscoring the need of experimental verification of structure following mutagenesis (Figure 3.6C).

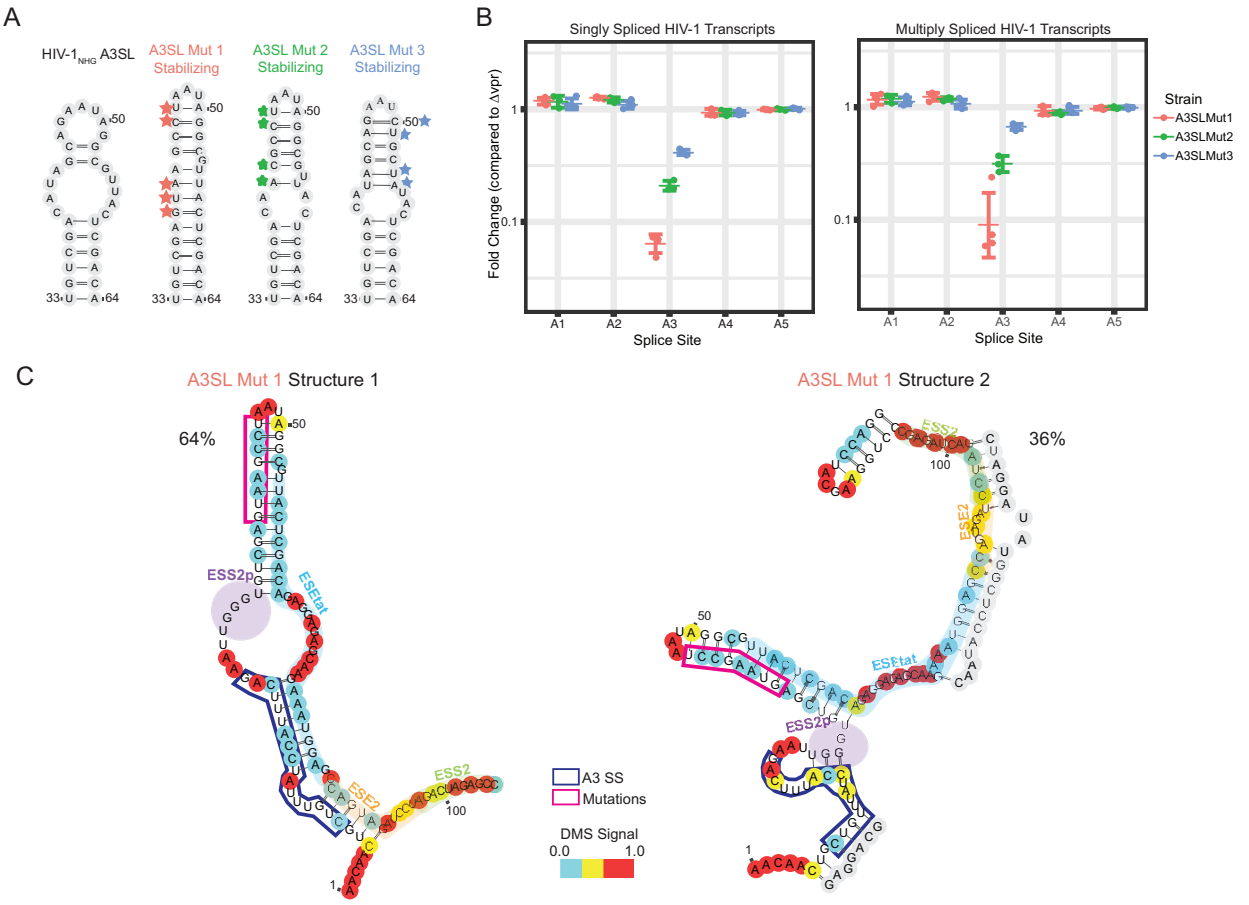

***Figure 3.6**: Mutations in the A3SL influence HIV-1 A3 splice acceptor site usage. A) Schematic showing the design rationale of the mutants. B) Splice junctions were measured by deep sequencing of cells transfected with the mutants. The splice usage at each acceptor site is reported as fold change compared to a reference strain, ΔvprHIV-1$_{NHG}$. C) DMS-modified RNA from HEK293t cells transfected with Mut1 was clustered by DREEM. Two clusters were identified and are shown with the location of the A3ss and known splice enhancer and silencer elements.*

In contrast, mutations in the same sequence region predicted to have little effect on the stability of A3SL, and therefore hypothesized to have little effect on splicing, increased A3ss usage relative to the parental strain (A3SLMut4, Figure 3.7A-B). To understand the origin for the increase in splicing, A3SL4 was probed and it was found that these mutations resulted in the formation of an unanticipated alternative structure in ~53% of molecules, demonstrating that thermodynamic predictions alone are incomplete. The unanticipated structure alters the accessibility of multiple nearby protein binding sites (Figure 3.7C). To further test the inhibitory role of A3SL, a compensatory mutant to shift the population towards the A3SL in the sequence context of the A3SLMut4 was designed. Consistent with A3SL inhibiting splicing, the compensatory mutant (A3SLMut5) uses A3ss ~10-fold less frequently than ΔvprHIV-1_{NHG} (Figure 3.7B). Together, these results indicate that the intrinsic ability of RNA to form alternative structures can regulate splicing either by directly occluding U2AF binding sites or by modifying the accessibility of nearby splicing enhancer and silencer elements, the net effect resulting in up to ~100-fold change in HIV-1 *tat* transcript abundance. The percentage of the A3SL cluster for each mutant had an inverse relationship with the overall usage of the A3 splice acceptor site (Figure 3.7D).

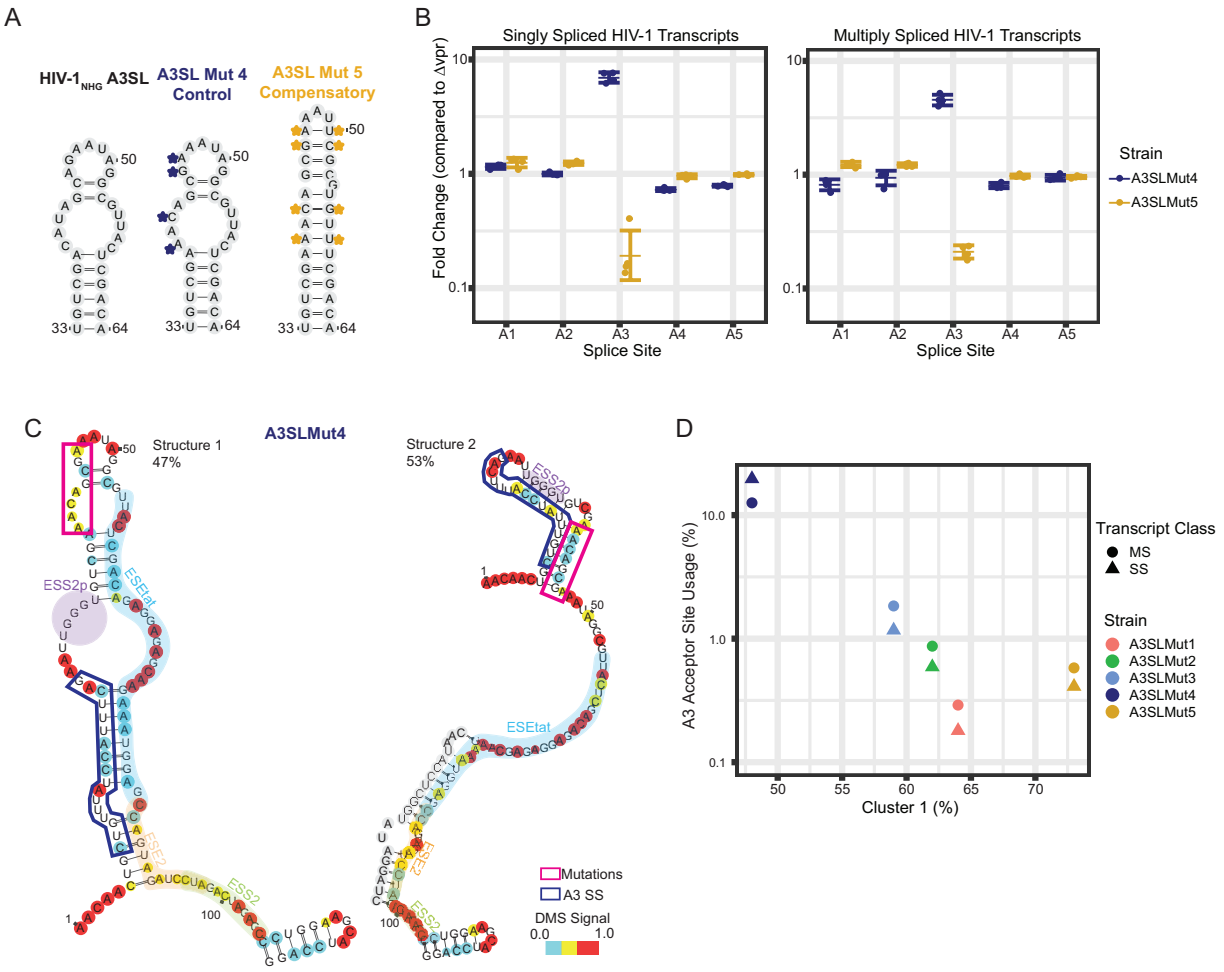

Figure 3.7*: Mutations in the A3SL have unpredicted effect on splicing due to alternative RNA structures. A) Design rationale of two additional A3SL mutants compared to HIV-1*$_{NHG}$ *as a reference. B) Splice junctions of A3SL Mut4-5 were measured by deep sequencing of cells transfected with the mutants. The splice usage at each acceptor site is reported as fold change compared to a reference strain, HIV-1*$_{NHG\Delta vpr}$*. C) DMS-modified RNA from HEK293t cells transfected with Mut4 was clustered by DREEM. D) A3 splice acceptor usage of all splice acceptor usage compared to percent cluster 1 (A3SL) as determined by DREEM for A3SL Mut1-5.*

Genome-wide alternative RNA structure analysis reveals widespread structural heterogeneity

To test whether formation of alternative structures is a general property of the HIV-1 RNA, a genome-wide DMS-MaPseq dataset from HEK293t cells transfected with HIV-1$_{NHG}$ was prepared. DREEM clustering was run on 80 nt overlapping windows spanning the entire genome

and a stringent Bayesian Information Criteria (BIC) test was applied to determine whether the data could be separated into two distinct structure signals[194]. Importantly, both the RRE and A3ss match the results obtained by specific RT-PCR.

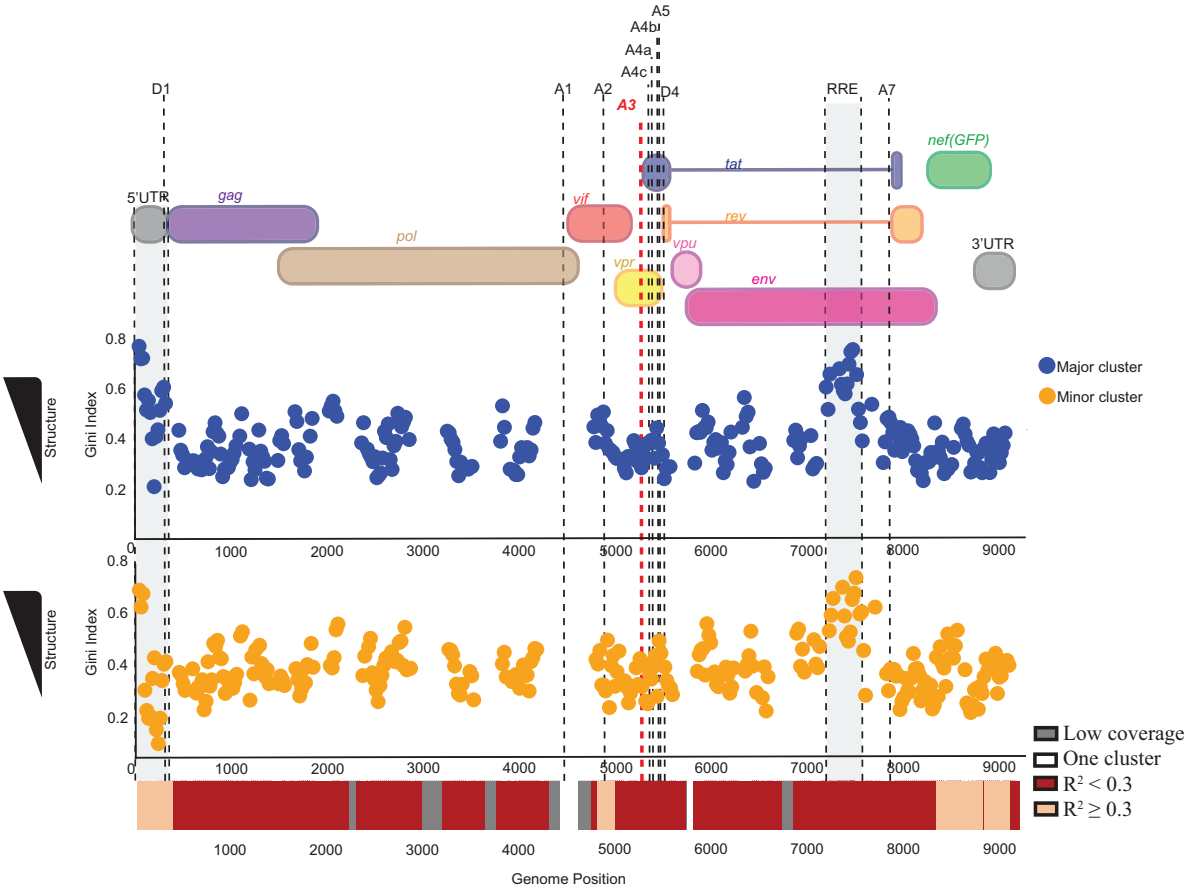

***Figure 3.8**: DREEM reveals RNA structural heterogeneity across the HIV-1 genome. On top is an overlay of the genetic organization of HIV-1. In the middle is a scatterplot where each dot represents an 80 nt window of the HIV-1 genome. Each window is analyzed based on reads from a whole-genome library from DMS-modified HEK293t cells transfected with HIV-1$_{NHG}$. The top scatterplot shows the higher percentage cluster as determined by DREEM. The bottom scatterplot is the less prevalent cluster. The Gini index was calculated for each cluster in each window. The bottom of the figure shows the results of correlation between the clusters of each window as shown in a heat map. Windows were there was insufficient coverage for clustering or in which only one cluster passed the BIC test are shown in grey and white respectively. Windows in which the two clusters had a Pearson's $R^2 \geq 0.3$ are shown in orange and windows with an $R^2 < 0.3$ are shown in red.*

Over 90% of windows with >100,000 sequencing reads coverage passed the BIC test for two clusters, indicating the presence of RNA structure heterogeneity across the entire HIV-1 genome. The extent of structure in each window was quantified using the Gini index metric, which measures the variability in reactivity of residues[141]. A Gini index close to zero indicates a relatively even distribution of DMS modifications, and occurs when RNA is unfolded or when RNA structure is highly heterogeneous. A Gini index close to one occurs when a subset of residues is strongly protected from DMS, and indicates a highly stable structure. The Pearson's correlation coefficient was computed for all windows that had alternative structures to measure how different the two structures were from each other. A low Pearson correlation (R^2<0.3) and low Gini index (<0.5) indicate that that relatively unstable, alternative structures form across the entire genome (Figure 3.8), including alternative conformations for a conserved structure[63] in the 4 kb *gag-pro-pol* region**,** which is present exclusively in unspliced transcripts. The smallest minor cluster that we observed was present at 20%, located in the *env* coding region.

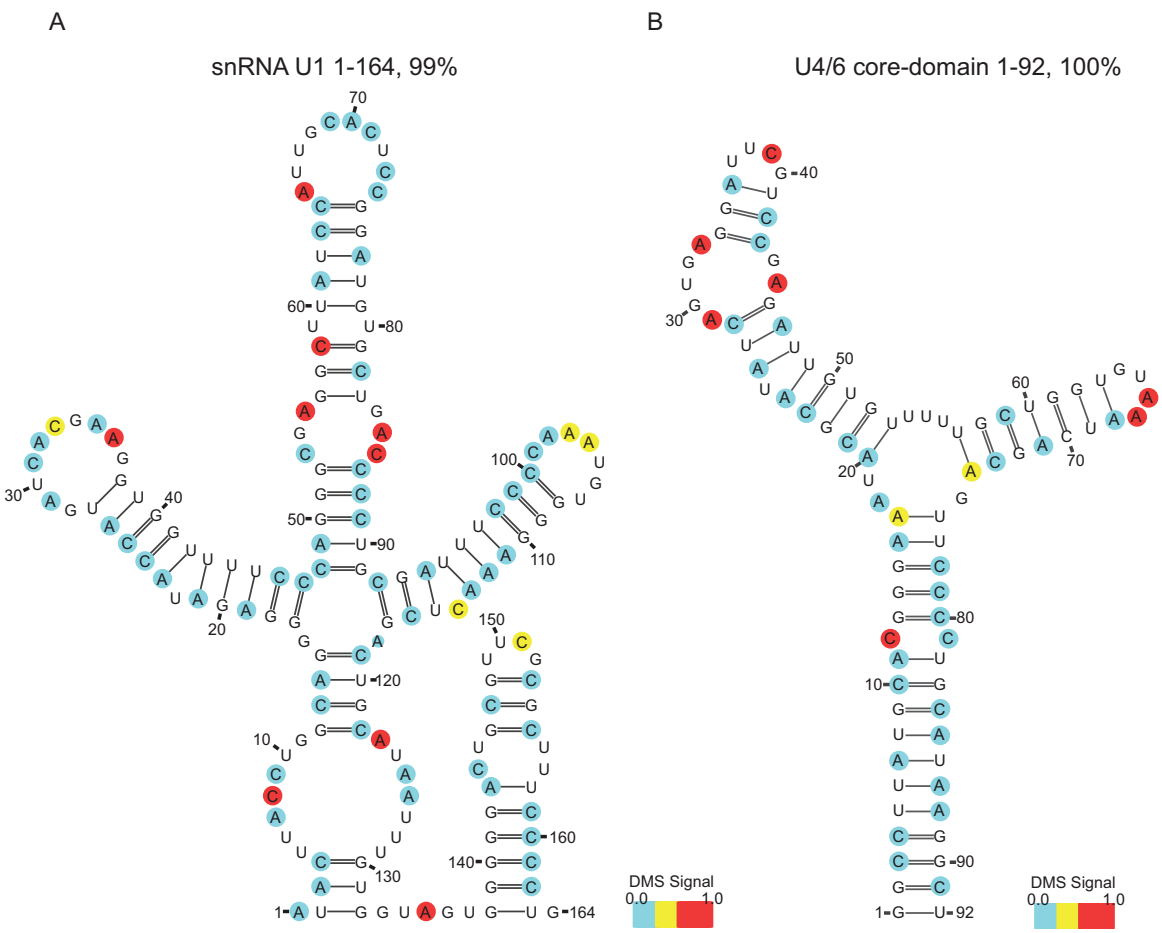

Figure 3.9*: snRNA U1 and U4/6 core-domain RNA structure predictions. A) The snRNA U1 from the whole-genome library of HEK293t cells transfected with HIV-1$_{NHG}$ was clustered by DREEM. One main cluster was detected at 99%. B) The U4/6 core-domain was* in vitro *transcribed, folded and DMS-modified. One cluster passed the BIC test after DREEM clustering.*

The widespread alternative structure of HIV-1 genome stood in contrast to snRNA U1 from the same dataset and U4/6 core-domain RNA probed *in vitro* (Figure 3.9). The snRNA U1 was found to have a cluster of 99% and U4/6 never had a second cluster pass the BIC. Both of these RNAs have stable structures determined by X-ray crystallography[195] and NMR[196] respectively. As a further control against over-clustering, simulated reads were generated based on the HIV-1 population average DMS-signal with no relationship between mutations. No windows of

simulated data passed the BIC test for two clusters. The whole-genome dataset was used to identify previously validated structures such as TAR[43,59,179], which was detected in one conformation (Figure 3.10A). Interestingly, RNA structural heterogeneity was detected at most splice sites including A4a-c, and A5 (Figure 3.10B). Together, these results suggest splice site occlusion as a general mechanism for HIV-1 to tune alternative splicing.

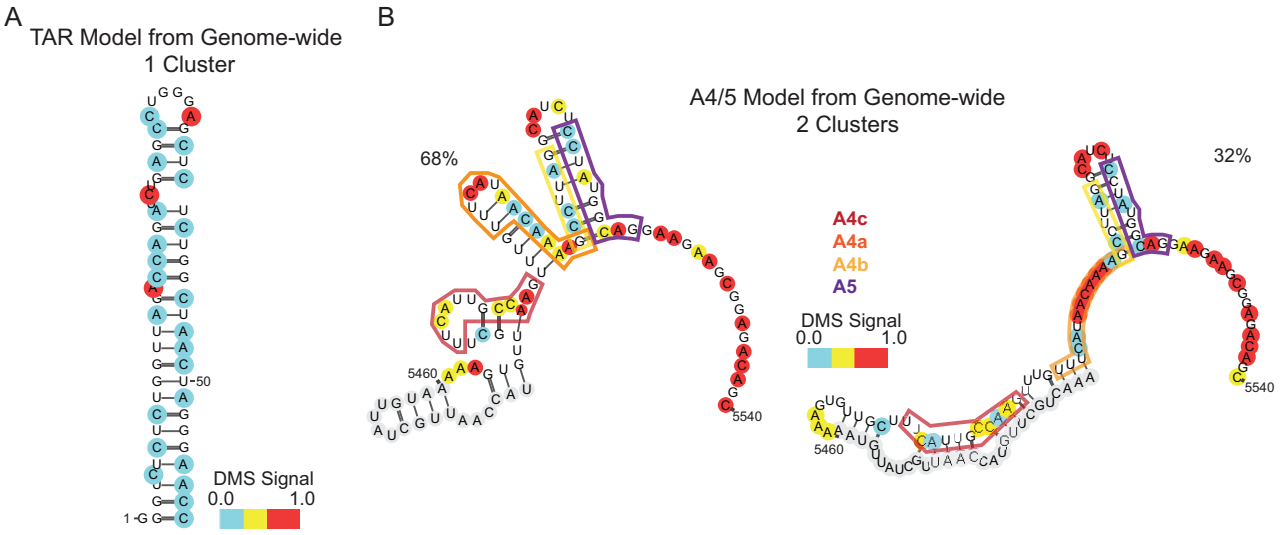

***Figure 3.10**: Genome-wide structure predictions for TAR and A4/5 splice acceptor site. A) TAR structure prediction from genome-wide library generated from HEK293t cells transfected with HIV-1$_{NHG}$. Clustering was done directly on the TAR region, and only one cluster passed the BIC. B) Structure prediction of A4/5 splice acceptor sites from whole-genome library. Two clusters passed the BIC test.*

3.4 Discussion

The validation experiments on DREEM using DMS-MaPseq data showed that the algorithm works for a variety of biological systems. The riboSNitch experiment showed that DREEM can resolve the known mixing ratio of two closely related RNA structures. The *add* experiment validated that DREEM can pick up RNA conformation changes induced by a biological stimulus. For the add experiment, DREEM was run without being constrained to a maximum of two clusters. For this example, up to four alternative structures that have been seen in previous studies were found. The experiments with the HIV-1 RRE showed that DREEM is able to make high-quality structure predictions from *in vitro* transcribed, folded and DMS-modified RNA constructs as well as from biological samples, namely RNA in cells and in virions.

The HIV-1 RRE was found to be consistent in different folding environments. This result indicated that the folding of the RRE is driven largely by thermodynamics and less by protein binding. Consistent with previous literature, our data imply that the conformational change induced by Rev binding is explained by tertiary structure changes but not secondary structure. We hypothesize that the stability of the RRE structure is a product of the extensive, highly self-complimentary helix that isolates the multi-stem region from the rest of the genomic RNA. This long helix reduces the possible folding conformations and helps the structure to fold consistently into a few conformations, which are able to be deconvoluted by DREEM. A similar alternative RNA structure at the end of the longest continuous helix in the HIV-1 genome was found to also have highly reproducible signal for DREEM.

In contrast to the RRE, the A3 splice acceptor site show differences when comparing intracellular structure to *in vitro* folded structure. Importantly, the A3SL formed in both conditions. The A3 splice site structure is therefore more dependent on protein binding for structure and not driven by thermodynamics alone. Despite the fact that the structure is less stable and more environment-dependent, the A3SL still has a biological function. When viewing the whole-genome clustering windows, there are many windows that match the description of the A3 splice site, i.e. low Gini index and low R^2 between clusters, whereas windows with the high Gini index of the RRE are rare. Taken together, these results indicate that biologically relevant RNA structures can occur throughout the whole genome, not just limited to the windows like the RRE with particularly high stability.

Chapter 4- CaptureSeq after reversal of HIV-1 latency identifies determinants of reactivation

Phillip Tomezsko[1,2,3], Kendyll Coxen[2], Heather Corry[2], Stephanie Banning[2], Olivia Roberts-Sano[2], Qianjing He[2], Silvi Rouskin[3], Daniel Kuritzkes[2,4], Athe Tsibris[2,4]

1. Program in Virology, Harvard Medical School, Boston, MA, USA

2. Brigham and Women's Hospital, Boston, MA, USA

3. Whitehead Institute, Cambridge, MA, USA

4. Harvard Medical School, Boston, MA, USA

Conceptualization: P.T, A.T. Methodology: P.T, S.R., D.K., A.T. Validation: P.T. Formal analysis: P.T., Investigation: P.T., K.C., H.C., S.B., O.R., Q.H. Resources: A.T. Data curation: P.T. Writing-original draft preparation: P.T. Writing-review and editing: P.T., S.R., D.K., A.T. Visualization: P.T. Supervision: D.K., A.T. Funding acquisition: S.R., D.K., A.T.

This manuscript is in preparation.

Abstract

HIV-1 reactivation of latent proviruses has been proposed as part of a therapeutic strategy to eliminate the viral reservoir. Despite some promise in preclinical studies, agents that target transcriptional repression of the latent proviruses have failed to reduce the size of the reservoir in clinical trials. A more comprehensive understanding of the host factors involved in successful reactivation from latency is critical for the use of novel latency reversing agents (LRA) or more potent combinations of current LRAs. In order to quantify HIV-1 RNA from latently infected cells treated with a diverse panel of LRAs, we developed a CaptureSeq approach with probes designed using participant-specific proviral sequences. Based on differential expression analysis of human genes, basic leucine zipper transcription factors were upregulated in all stimulation conditions. CaptureSeq proportionally enriched mRNA transcripts from HIV-1 and from a set of control genes with an average of ~500-fold enrichment. The enriched RNAseq data were used to determine the splicing ratio of HIV-1 transcripts and no relationship between HIV-1 expression and splicing was observed. This technique and the insights generated from this study can be used to analyze novel LRA and LRA combinations in order to reactivate latent proviruses more specifically.

4.1 Introduction

HIV-1 seeds a reservoir early in infection that can be reactivated after even decades on suppressive antiretroviral therapy[77,78,81]. Most infected individuals rebound within weeks of stopping antiretroviral therapy (ART)[80]. Although early ART initiation significantly limits the size of the reservoir, reactivation upon ART cessation occurs even if ART was started soon after infection[82,197]. All subsets of $CD4^+$ T cells harbor latently infected cells; however, the largest known constituent of the latent reservoir is resting, memory $CD4^+$ T cells[83,87,198]. $CD4^+$ T cells in tissue compartments can also be latently infected, and latently infected cells in certain tissues can have higher transcriptional activity at baseline compared to circulation $CD4^+$ T cells in the blood[199]. Memory T cells are long-lived; clonal expansion of latently infected cells maintains the reservoir for long periods[98,99,105,200-203]. Virus from reactivated clonally expanded cells is the most likely cause of viral rebound after treatment interruption[204]. Clonal expansion and contraction of particular lineages of latently infected T cells occurs over the course of the infected individual's life time, although cells harboring intact proviruses decay over time[101,102,108].

The HIV-1 provirus is transcriptionally repressed through multiple mechanisms in latently infected cells. The provirus associates with nucleosomes upon integration[88]. In resting $CD4^+$ T cells, the proviral nucleosomes are modified with repressive epigenetic marks such as methylation[90,91]. Activating histone modifications, such as acetylation, are removed[92]. Key transcription factors including NF-kB and NFAT are localized in the cytoplasm in resting $CD4^+$ T cells[93]. The RNA elongation factor P-TEFb, which is recruited to the HIV-1 LTR by Tat and TAR during production infection, is sequestered in resting $CD4^+$ T cells[94]. In the natural course

of infection, T cell activation serves as the strongest stimulus for reactivation, as all these mechanisms are linked to the T cell receptor and co-stimulatory signaling cascade. Strength of TCR stimulation correlates with inducibility of the HIV-1 provirus[95].

Latency reversing agents (LRA) have been used both *in vitro* and in clinical trials to reactivate latent HIV-1 proviruses in order to expose them to killing by the immune system, theoretically in combination with an agent to enhance immune recognition and killing. LRAs have been tested that target each of the aforementioned mechanisms of transcriptional regulation. Histone deacetylase inhibitors (HDACi) have been tested to reverse epigenetic silencing of latent proviruses[205]. Despite some promising preclinical studies, clinical administration of HDACi's romidepsin, panobinostat and vorinostat have resulted in short-lived increases in cell-associated HIV-1 RNA, but no increase in decay of the reservoir[110,124,126,205]. PKC agonists, including bryostatin-1 and ingenols, target a key node of the T cell activation pathway and induce NF-kB signaling[112,113]. Although bryostatin-1 was safe in clinical trials at low doses, it was not effective in reducing the latent reservoir[116]. Immunomodulatory agents are being investigated to reactivate latently infected cells through more physiological stimuli without causing immune pathology. These immunomodulatory LRAs include TLR agonist, IL-15 and IL-15 superagonist[117-119,121,122]. Based on the lack of efficacy from single agents, combinations of LRAs with different mechanisms of action have been explored[133,134]. However, the first reported study of a combination HDACi along with a therapeutic vaccine showed increase in HIV-1 RNA without a decrease in the viral reservoir[206].

Beyond developing new LRAs and combinations, renewed interest has been shown towards deciphering the determinants of successful reactivation, beyond what is currently known. Regulation of a diverse array human co-factors is crucial for reactivation of the HIV provirus. Three landmark single cell (sc) RNAseq papers, two in primary latency models and one from using reactivated cells from ART suppressed study participants with HIV-1, provided insight into the human genes and regulatory patterns that impact reactivation[207-209]. One of the most important insights from the study in latently infected cells isolated from primary cells is that there is an alternative transcription factor (TF) profile in cells that reactivate compared to cells that do not reactivate[209]. Studies done in primary latency models showed that reactivation corresponds with T-cell activation but not proliferation[207] and that populations of cells exist in two states, one that is susceptible to reactivation and one that is not[208]. This model could explain the heterogeneity between cells in response to LRAs.

Differential TF profiles are noteworthy because the HIV-1 promoter is quite complex. The promoter shares similarities with several genes encoding cytokines, such as TNFα and IL-6. However, HIV-1 transcription can be promoted by numerous pathways. HIV-1 transcription can be promoted via NF-kB independent or alternate NF-kB pathways, as well as traditional NF-kB[210,211]. The HIV-1 promoter also has AP-1 binding sites that are important for the establishment and reactivation of latency[212,213].

We developed a probe-based CaptureSeq technique for quantifying HIV-1 transcription using RNAseq from ART-suppressed study participants. The probes were designed based on primary HIV-1 sequences from the participants. Primary cells were stimulated *ex vivo* by a panel of

LRAs and mRNA was isolated for analysis by RNAseq and CaptureSeq. We identified the transcription factor landscape in each stimulation condition and compared the available TFs to binding sites in the HIV-1 promoter. We found that the HIV-1 promoter is able to bind to a diverse array of TFs that are differentially expressed after stimulation with a variety of LRAs. Members of the basic leucine (bZIP) family may have broader binding capacity than previously known based on the number of half-sites in the HIV-1 promoter, compensating for conditions in which NF-kB and NFAT are not available. After enrichment, we observed RNA coverage across the HIV-1 genome and compared the splicing ratio to HIV-1 expression. We observed no relationship between splicing and HIV-1 RNA expression.

4.2 Methods and Materials

Study participants and samples

Four study participants were enrolled from the HIV Eradication after Latency (HEAL) cohort. Each participant was virally suppressed on ART. Participants donated large volume leukapheresis. Peripheral blood mononuclear cells (PBMCs) were isolated by Ficoll-Histopaque density centrifugation. PBMCs were cryopreserved at a concentration of 50 million/mL.

Sample Treatment and RNA extraction

Non-naïve, resting $CD4^+$ T cells (nn-rCD4) were isolated from fresh or frozen cells using a custom immunomagnetic isolation kit (Stem Cell, USA). The kit contained the labeled antibodies against the following to negatively select nn-rCD4: CD8, CD14, CD16, CD19, CD20, CD25, CD36, CD56, CD66b, CD69, CD123, HLA-DR, GlyA and CD45RA. nn-rCD4 cells were cultured at a concentration of ~1 million/mL in RPMI media supplemented with 10% fetal bovine serum and 50 U/mL penicillin-streptomycin (R10 media).

15 million nn-rCD4s were stimulated for 24 hours with the following conditions in 15 mL of R10 media: 0.1% DMSO (unstimulated), 50 ng/mL phorbol 12-myristate 13-acetate and 1μM ionomycin (PMA-iono; Millipore-Sigma), 20 nM romidepsin (RMD; Selleckchem, USA), 10 nM Bryostatin-1 (Bryo; Millipore-Sigma), 1.5 nM IL-15 (R&D Systems, USA), 20 nM RMD and 1 nM Bryo (RB). After 24 hours, the cells were centrifuged at 1000xg for 10 minutes, supernatant was removed and the pellets were resuspended in 1 mL Trizol (ThermoFisher Scientific). 250 μL of chloroform was added to the Trizol, vortexed and centrifuged for 15 mins at 15,000xg 4ºC. The aqueous phase was transferred to a new tube and an equal volume of

isopropanol, 100 µL of 3 M sodium acetate and 3 µL of glycoblue (ThermoFisher Scientific) were added. The samples were frozen on dry ice. The samples were centrifuged for 45 mins at 15,000xg 4ºC. The supernatant was removed, the pellet was washed with ice-cold 70% ethanol and resuspended in 20 µL of nuclease-free water. Concentration was measured by Nanodrop.

Library Generation

In order to generate RNAseq libraries, 5 µg of extracted RNA for each condition was used for library generation. Poly-adenylated RNA was isolated using the Oligo(dt) 25 poly-A selection kit (ThermoFisher Scientific). The poly-adenylated RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of all fragments and eluted in 9 µL of nuclease-free water (Zymo Research, USA). The RNA was fragmented using the RNA Fragmentation kit (ThermoFisher Scientific) with a fragmentation step of 90 seconds at 70ºC. The RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of all fragments and eluted in 6.5 µL of nuclease-free water. 1 µL of CutSmart buffer (New England Biolabs), 1.5 µL of Shrimp Alkaline Phosphatase (New England Biolabs) and 1 µL of RNaseOUT (ThermoFisher Scientific) were added and incubated at 37ºC for 1 hour to dephosphorylate the RNA. 6 µL of 50% PEG-800 (New England Biolabs), 2.2 µL of 10x T4 RNA Ligase buffer (New England Biolabs), 2 µL of T4 RNA Ligase, truncated KQ (England Biolabs) and 1 µL of linker were added to the reaction and incubated for 18 hours at 22ºC. The RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery of all fragments and eluted in 15 µL of nuclease-free water. Excess linker was degraded by adding 2 µL of 10x RecJ buffer (Lucigen), 1 µL of RecJ exonuclease (Lucigen), 1 µL of 5'Deadenylase (New England Biolabs) and 1 µL of RNaseOUT, then

incubating for 1 hour at 30ºC. The RNA was purified with RNA Clean and Concentrator -5, following the manufacturer's instructions for recovery fragments > 200 nt and eluted in 11 µL of nuclease-free water. For reverse transcription, 1 µL of RT primer, 1 µL of 0.1M DTT, 4 µL of 5x First Strand buffer, 1 µL of dNTP, 1 µL of RNaseOUT and 1 µL of T-GIRT III were added and the sample was incubated for 2 hours at 65ºC. RNA was degraded by adding 1 µL of 4 N sodium hydroxide and incubating at 95ºC for 3 mins. The RT product was mixed with an equal volume of 2x Novex TBE-Urea sample buffer (ThermoFisher Scientific) and run on a pre-cast 10% TBE-Urea gel (ThermoFisher Scientific) and the ~300-400 nt product was extracted. The purified ssDNA was circularized using the CircLigase ssDNA Ligase kit (Lucigen). 2 uL of the circularized product was used for PCR using Phusion and an indexed forward primer for multiplexing. The sample was run for a maximum of 14 cycles. Following PCR, the product was run on a pre-cast 8% TBE gel and the ~250-350nt product was gel extracted. The final PCR product was quantified by Bioanalyzer (Agilent). The libraries were sequenced on an Illumina NextSeq to obtain 75x75 paired-end reads.

HIV-1 Single Genome Amplification

For each participant, 10 million PBMCs were thawed and resuspended in R10 media. DNA was extracted using the AllPrep DNA/RNA kit following manufacturer's instructions (Qiagen, Germany). DNA was eluted into a final volume of 200 µL of TE buffer and concentration was measured by Nanodrop. DNA was diluted 1:2, 1:4, 1:8 and 1:10 in order to determine the optimal dilution for single genome amplification (SGA). Twelve reactions of each dilution were amplified by nested PCR using Platinum Taq DNA polymerase (ThermoFisher Scientific, USA). For primers, see section below. The PCR products were run on 1% agarose gels in order to

determine the number of positive reactions. The lowest dilution that yielded <30% positive was used for sequence generation. 96 nested PCR reactions were set up with the selected dilution. PCR product was visualized on agarose gel and positive reactions with a product of ~8 kb were sequenced. HIV-1 sequences that were over 8 kb and were not hypermutated as determined by Hypermut 2.0 (https://www.hiv.lanl.gov/content/sequence/HYPERMUT/background.html) were included in probe design.

Probe Enrichment

HIV-1 and human control gene sequences were used to design tiling biotinylated oligos using the myBaits custom probe set (Arbor Biosciences, USA). Three 8-cycle PCR reactions using the circularized product from the library generation as a template were run using Phusion for each sample. After the PCR was complete, the three reactions were combined. The DNA was purified using DNA Clean and Concentrator (Zymo Research, USA), and eluted in a final volume of 7 μL of nuclease-free water. The hybridization reaction with custom baits was mixed into a final volume of 18 μL according to manufacturer's specifications and incubated for 10 minutes at 60ºC then room temperature for 5 minutes. The blocking oligos and 7 μL of PCR product were mixed to a final volume of 12 μL and incubated for 5 minutes at 95ºC. Both the blocker and hybridization reactions were incubated at 65ºC for 5 minutes. The blocker and hybridization reactions were mixed and incubated for 12-16 hours at 65ºC. Magnetic streptavidin beads were incubated for 65ºC for 2 minutes in a volume of 70 μL binding buffer. The beads and samples were mixed and incubated together for 5 minutes at 65ºC. The samples were alternatively washed with the provided wash buffer on the magnetic stand and incubated for 5 minutes at 65ºC three times. After the final wash, the samples were placed on the magnetic stand, the supernatant

was removed and resuspended in 30 µL of 10 mM Tris solution with 0.05% tween. The samples were incubated for 5 minutes at 95°C, transferred to the magnetic stand and the elutant was immediately transferred to a new tube. The DNA was purified using DNA Clean and Concentrator and eluted in a final volume of 13 µL of nuclease-free water. The DNA was amplified by PCR using Phusion and an indexed forward primer for 12 cycles. The PCR product was resolved on a pre-cast 8% TBE gel and the correct sized product (~250-300 nt) was gel extracted and purified into a final volume of 10 µL of nuclease-free water. The concentration of the final enriched libraries was determined by Bioanalyzer. The libraries were sequenced on an Illumina NextSeq to obtain 75x75 paired-end reads.

Bioinformatic Analysis

Low quality reads were removed using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/commandline.html). Paired-end reads were de-duplicated by unique molecular identifier using UMI_tools (https://github.com/CGATOxford/UMI-tools). A primary human transcriptome was obtained from Gencode (https://www.gencodegenes.org/human/), build GRCh38.12. Consensus HIV-1 sequences for each participant were obtained using Geneious. The de-duplicated reads were aligned separately to the human transcriptome and the HIV-1 using Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). Human and HIV-1 alignments were combined and converted to counts per gene or HIV-1 region using a custom python script. The counts were used for differential gene expression analysis using DEseq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html). Gene ontology was performed on differentially expressed gene sets using GOseq

(https://bioconductor.org/packages/release/bioc/html/goseq.html). Images were generated on R version 3.5.1 using ggplot2 (https://ggplot2.tidyverse.org/). Heatmaps were generated using the heatmap.2 function in gplots (https://cran.r-project.org/web/packages/gplots/index.html). Enriched HIV-1 aligned to consensus sequences were visualized using IGV (http://software.broadinstitute.org/software/igv/) Schematic in figure 1 was created with Biorender (https://app.biorender.com/).

HIV-1 Baseline Reservoir Measurement

For each participant, 75 million PBMCs were thawed for each participant. Non-naïve, resting $CD4^+$ T cells were isolated using the custom immunomagnetic separation kit from 70 million PBMCs. DNA and RNA were extracted from the remaining 5 million PBMC and up to 5 million non-naïve, resting $CD4^+$ T cells using the AllPrep DNA/RNA kit according to manufacturer's specifications. The extracted DNA and RNA were used for qPCR and RT-qPCR respectively to quantify HIV-1 cell-associated copies per million cells. HIV-1 caRNA copies were measured using 5'LTR-Gag specific primers and a FAM-labeled probe (ThermoFisher Scientific, USA) with TaqMan Fast Virus 1-Step Master Mix (Applied Biosystems). HIV-1 caDNA copies were measured using the same 5'LTR-Gag specific primers and a FAM-labeled probe (Millipore-Sigma, USA) with TaqMan Universal PCR Master Mix (Applied Biosystems). The cells in each sample were quantified by running the extracted DNA with CCR5 specific primers and FAM-labeled probe (Millipore-Sigma, USA) with TaqMan Universal PCR Master Mix (Applied Biosystems). The PCR was run on the ABI 7300. This experiment was run in triplicate and reported as the mean of the triplicates. Values for PBMC and nn-rCD4 were compared by two-tailed, paired T-test.

HIV-1 Reservoir Measurement after Latency Reversal

In order to quantify the latent reservoir to compare to RNAseq, 250-300 million PBMC from each participant were thawed. Non-naïve, resting $CD4^+$ T cells were isolated using the custom immunomagnetic separation kit from all thawed cells. Five million cells per condition were stimulated for 24 hours with the same six conditions as above. DNA and RNA were extracted from isolated nn-rCD4 cells using the AllPrep DNA/RNA kit according to manufacturer's instructions. The same primers, probes and RT-qPCR and qPCR conditions were used as for the baseline reservoir measurements.

Statistical Analysis

The fold-enrichment of baseline measures of nn-$rCD4^+$ T cells were tested with a one-sample T-test. Differential gene expression and PCA analysis was performed using DESeq2 with a $p > 0.01$ threshold. GO analysis was performed with GOseq using a $p > 0.05$ threshold for significance. ecdf curves and comparison of DE gene expression was performed by Wilcoxon Rank Sum test with a Benjamini-Hochberg correction for multiple comparisons. All correlations were measured for significance using Pearson's correlation coefficient. All statistical tests were performed on excel or R.

Primers and Oligos

(All oligos ordered from IDT, USA unless noted otherwise).

Library generation linker:

/5rApp/TCNNNNNNNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGA/3ddC/

Library generation RT primer:

/5Phos/AGATCGGAAGAGCACACGTCTGAACTCCAG/iSp18/TCTTTCCCTACACGACGC

TCTTCCGATCT

Library generation forward PCR primer:

CAAGCAGAAGACGGCATACGAGAT**XXXXXX**GTGACTGGAGTTCAGACGTGTGCTC

Library generation reverse PCR primer:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC

SGA PCR 1 Forward: AAATCTCTAGCAGTGGCGCCCGAACAG

SGA PCR 1 Reverse: TGAGGGATCTCTAGTTACCAGAGTC

SGA PCR 2 Forward: GCGCCCGAACAGGGACYTGAAARCGAAAG

SGA PCR 2 Reverse: GCACTCAAGGCAAGCTTTATTGAGGCTTA

HIV qPCR Forward: TACTGACGCTCTCGCACC

HIV qPCR Reverse: TCTCGACGCAGGACTCG

HIV qPCR Probe: 5' FAM-CTCTCTCCTTCTAGCCTC-MGB 3' (ThermoFisher Scientific)

CCR5 qPCR Forward: ATGATTCCTGGGAGAGACGC

CCR5 qPCR Reverse: AGCCAGGACGGTCACCTT

CCR5 qPCR Probe: 5' FAM-AACACAGCCACCACCCAAGTGATCA-BHQ (Millipore-Sigma)

4.3 Results

Experimental setup and baseline measurements of latent reservoir

Large-volume blood draws were obtained from four study participants with HIV-1 from the HIV-1 eradication after reversal of latency (HEAL) cohort. Because latently infected cells are so rare within ART-suppressed individuals, an enrichment strategy was developed in order to achieve coverage of RNAseq reads across the HIV-1 genome (Fig 4.1A). In order to maximize the population of latently infected cells used for the RNAseq library generation, non-naïve, resting $CD4^+$ T cells (nn-r$CD4^+$ T cells) were isolated from PBMCs by negative selection with magnetically labeled antibody-bead conjugates. Fifteen million nn-r$CD4^+$ T cells were treated with one of the following conditions for 24 hours: unstimulated (0.1% DMSO), 50 ng/mL phorbol 12-myristate 13-acetate and 1μM ionomycin (PMA-iono), 20 nM romidepsin (RMD), 10 nM bryostatin-1 (Bryo), 1.5 nM IL-15 (IL-15), 20 nM romidepsin and 1 nM bryostatin-1 (RB). After stimulation, RNA was extracted for RNAseq library generation. The unenriched libraries were sequenced and used for differential host gene expression analysis. The cDNA for HIV-1 and ten control host genes were enriched using tiling biotinylated probes. The host genes were selected to quantify a range of expression from medium to low in order to compare the unenriched libraries to the enriched libraries. The enriched libraries were then sequenced and HIV-1 expression was quantified. The size of the latent reservoir in nn-r$CD4^+$ T cells was compared to PBMC for each of the four participants. qPCR of cell-associated (ca)RNA and DNA from the proximal *gag* region was used to measure the reservoir. A 3.2-fold increase in HIV-1 DNA was observed in nn-r$CD4^+$ T cells compared to PBMC ($p = 0.017$) and a 3.8-fold increase in HIV-1 RNA was observed, though this change was not statistically significant ($p=0.34$) (Figure 4.1B).

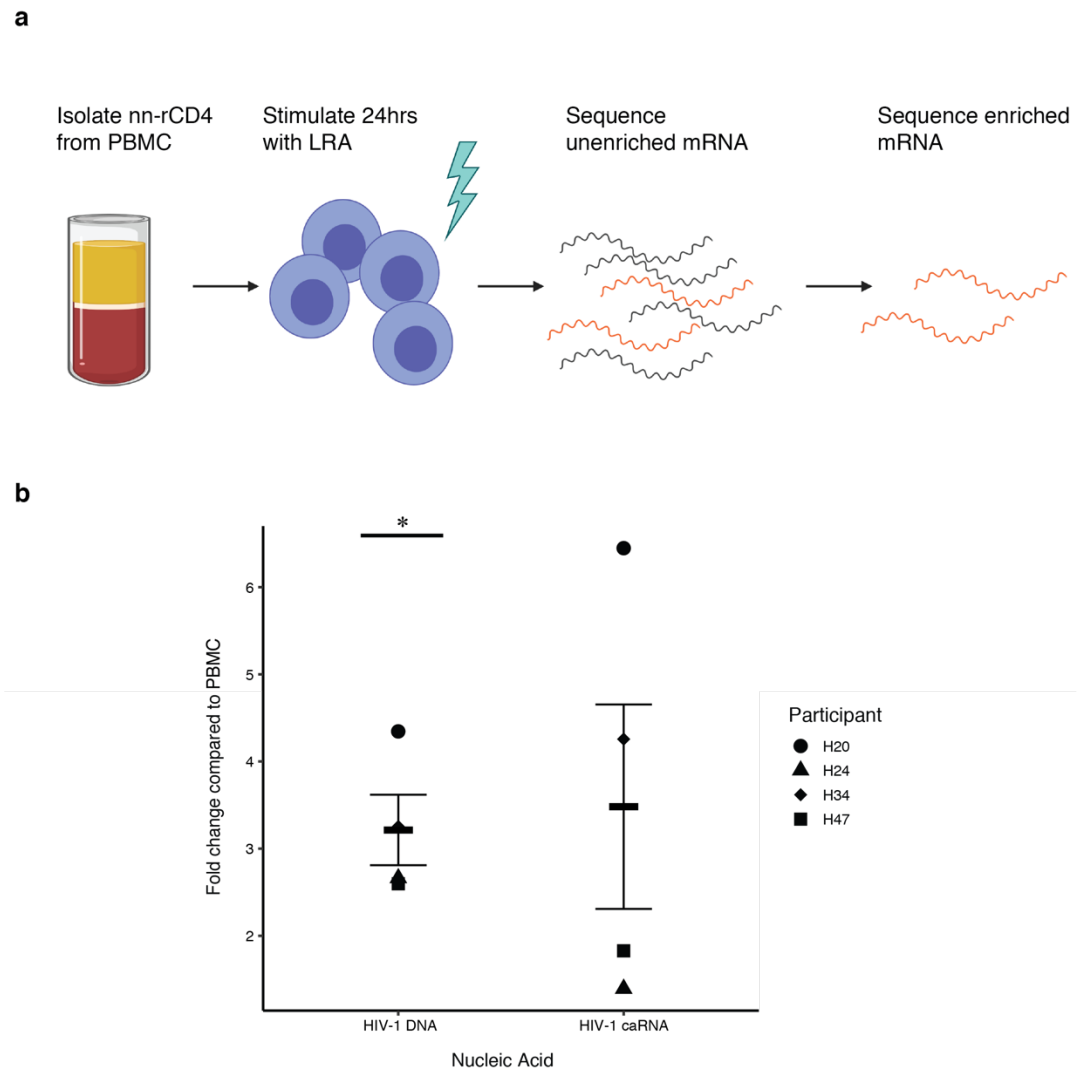

***Figure 4.1**: CaptureSeq for enrichment of HIV-1 from non-naïve, resting CD4⁺ T cells. a) Schematic of isolation, stimulation and mRNA isolation for unenriched and enriched library generation. b) Fold change of HIV-1 DNA and caRNA levels compared to PBMC for nn-rCD4⁺ T cells as measured by qPCR (n=4 participants). Individual values are plotted with the mean and error bars showing s.e. * denotes $p < 0.05$, as determined by a one sample T-test compared to 1 (no fold-change).*

Host transcriptomic analysis of stimulated non-naïve, resting CD4$^+$ T cells from HIV-1+ participants

RNA was extracted from nn-rCD4$^+$ T cells that were stimulated with the LRA panel; mRNA was purified by selection for poly-adenylated tails and used for library generation. Unenriched cDNA was amplified and sequenced. After sequencing, the reads were filtered for quality and aligned to a primary human transcriptome from Gencode (build GRCh38.12) using Bowtie2. Reads were also aligned to a consensus HIV-1 sequence generated for each of the four participants from near-full length sequences generated by single genome amplification. A neighbor-joining tree was generated to quantify sequence diversity within participants and also ensure that the sequences were not cross-contaminated. Between six to twelve near-full length, non-hypermutated sequences were generated for each participant.

After alignment to both the human transcriptome and the HIV-1 genome, the alignments were converted to counts per gene using a custom python script. The unnormalized counts were used for differential gene analysis using the R package DESeq2[214]. DESeq2 creates a negative binomial model that accounts for both shot noise and inter-participant variability. A principal components analysis (PCA) was performed on the RNAseq dataset (Figure 4.2A). The PCA plot showed distinct clustering of the six stimulation conditions for the four participants. Compared to unstimulated cells, stimulation with PMA-iono was the furthest on the primary axis. By contrast, the RMD stimulation condition was furthest on the secondary axis. Differentially expressed and upregulated genes with a p value less than 0.01 and a $\log_2$-fold change greater than 2.0 were used for gene ontology analysis using GOseq. In order to gain specificity, differential gene sets were identified from sequential pairs along the primary and secondary axis and the overlapping GO

terms were analyzed. For example, in order to dissect the GO terms along the primary axis, the differentially expressed genes from PMA-iono/RB, RB/RMD, PMA-iono/Bryo and Bryo/Unstimulated were used and GO terms that were enriched in each condition were identified (Figure 4.2B). The top ten GO terms as determined by p value in PMA-iono/RB were plotted, along with the p value for each comparison. All ten terms are related to the immune response, as expected given the stimulation conditions. Of note, several terms are related to cytokine regulation and response.

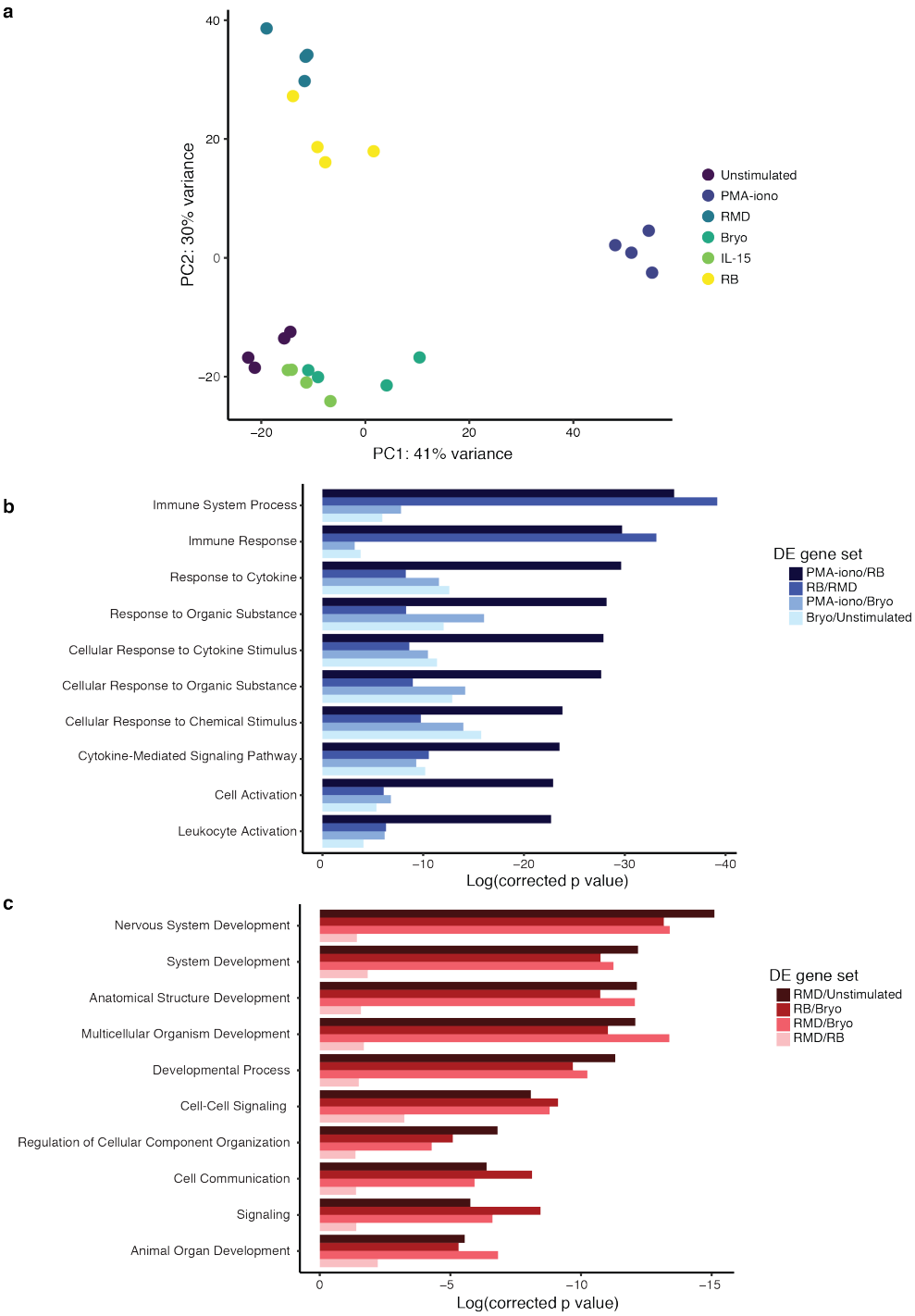

***Figure 4.2**: Host transcriptomic analysis after reversal of latency. a) Principal components analysis (PCA) of RNAseq reads from six stimulation conditions (n=4 participants). PCA plot generated by DESeq2. b) Gene ontology (GO) analysis displaying the top ten overlapping terms using differentially expressed genes from the following conditions (numerator/denominator): PMA-iono/RB, RB/RMD, PMA-iono/Bryo and Bryo/Unstimulated using a*

significance threshold of 0.01 and a log2 fold-change > 2. The order of GO terms was determined by the p value of the PMA-iono/RB condition. GO analysis was performed using GOseq, using a threshold of 0.05 corrected p value to determine significance. c) GO analysis displaying the top ten overlapping terms using differentially expressed genes from the following conditions: RMD/Unstimulated, RB/Bryo, RMD/Bryo and RMD/RB using a significance threshold of 0.01 and a log2 fold-change > 2. The order of GO terms was determined by the corrected p value of the RMD/Unstimulated condition. GO analysis was performed using GOseq, using a threshold of 0.05 corrected p value to determine significance.

Along the secondary axis, differentially expressed genes from RMD/Unstimulated, RMD/Bryo, RB/Bryo and R/RB were used for GOseq (Figure 4.2C). The shared top ten GO terms for this set did not seem to represent physiological stimulation. There were several terms related to the nervous system and development. The differentially expressed genes that contributed to these GO terms seemed to have lower expression values than for the other sets. This observation led to the hypothesis that stimulation with the HDACi romidepsin allowed for basal transcription to be increased of many disparate genes, largely representing an increase in noise. In order to test this hypothesis, we compared the mean expression of the differentially expressed genes in each condition compared to unstimulated (Figure 4.3A). We found that the expression was higher after stimulation with PMA-iono than each other condition with a median of 763.9 TPM. Stimulation with Bryo and IL-15 yielded median values of 305.4 TPM and 368.5 TPM, which were higher than RMD and RB stimulation with medians of 82.7 and 86.0 TPM. To compare the expression of each gene, an empirical cumulative distribution function curve (ecdf) was used. Comparing the most extreme sets, PMA-iono and RMD, showed that the curves were significantly different ($p < 2*10^{-16}$). PMA-iono stimulation had a higher number of genes at a low expression range (<1 TPM) and high expression range (>100 TPM). The RMD stimulation curve had more genes in the intermediate range (1-100 TPM) (Fig 4.3B). A similar relationship

was found for all conditions plotted on an ecdf plot. The PMA-iono distribution was distinct from all other conditions, Bryo and IL-15 distributions were distinct from all other conditions and unstimulated, RMD and RB stimulations had no difference between one another. Additionally, RMD and RB stimulation had the lowest ratio of differentially expressed genes to enriched GO terms, indicating that transcription was not happening in cohesive transcriptional units.

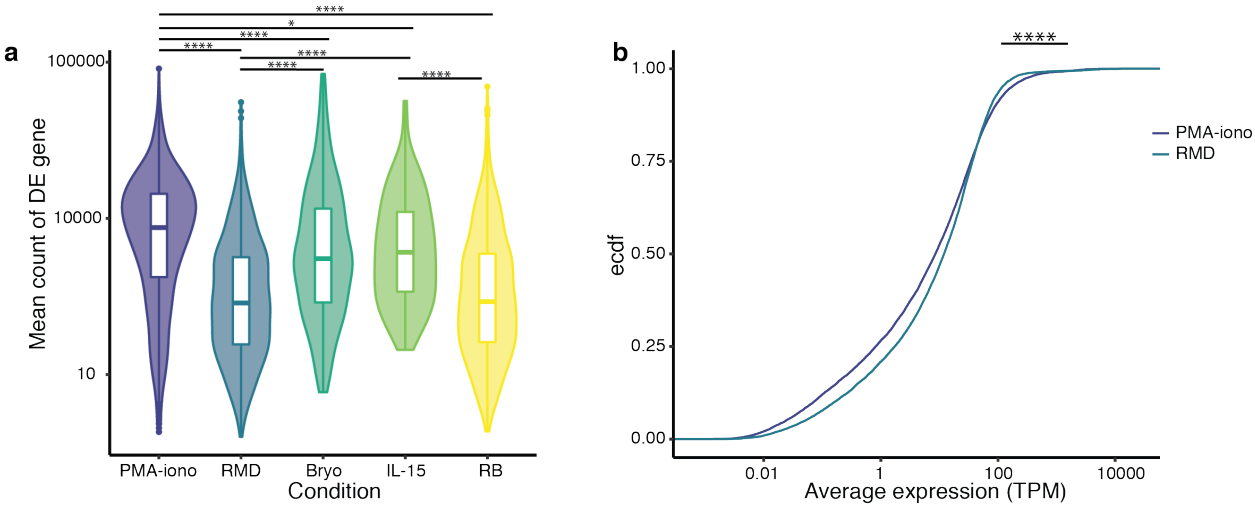

Figure 4.3**: Latency reversing agents create global transcription changes. a) The mean values of expression as calculated by DESeq2 for each upregulated gene compared to unstimulated is plotted for the LRA stimulations (n=4 participants). Upregulated genes were determined by DESeq using a significance threshold of 0.01 and a log2 fold-change > 2. The values are shown as a violin plot with median and interquartile boxplot plus outliers. * denotes p < 0.05, ** p < 0.01, *** p < 0.001, **** p < 0.0001 as determined by Wilcoxon Rank Sum test corrected for multiple comparisons by Benjamini-Hochberg (BH) procedure. b) An empirical cumulative distribution function describing the average expression as calculated by DESeq2 of all genes after stimulation with PMA/iono and RMD (n=4 participants). * denotes p < 0.0001 as determined by Wilcoxon Rank Sum test.*

GOseq analysis of PMA-iono, RB and Bryo stimulation revealed that factors that regulate response to and production of cytokines were differentially expressed. Therefore, the levels of cytokine expression in each group were compared (Figure 4.4A). An extensive but not exhaustive list of cytokines was grouped by family based on receptor identity[215,216]. Each point represents the average of the four participants for each cytokine. Cytokines that were found to be differentially expressed compared to unstimulated by DESeq2 are shown in red; cytokines that are not differentially expressed are shown in grey. PMA-iono stimulation had the greatest number differentially expressed cytokines (n=26) and RB stimulation had the least, with two out of 74 cytokines tested. The most differentially expressed cytokine families after PMA-iono stimulation are in the IL-2, IL-6 and TNF families.

We observed that many of the differentially expressed genes that contributed to the immune function GO terms were transcription factors (TF). The values of all human TFs were compared on a heatmap (Figure 4.4B)[217]. Interestingly, PMA-iono stimulation has the lowest overall expression of all transcription factors, but had a number of highly-upregulated TFs. PMA-iono stimulation also had the highest number of differentially expressed TFs (Figure 4.4C). Many of the upregulated TF after PMA-iono stimulation belonged to NF-κB and bZIP families, both important families in the T cell activation signaling pathway. The bZIP proteins were particularly interesting because they are thought to bind to the AP-1 and CREB sites of the HIV-1 promoter. However, the breadth and diversity of bZIP binding to the HIV-1 promoter has not been fully explored. When comparing the differentially expressed TF in each LRA stimulation group to the unstimulated gourp, we observed that there was substantial overlap in the bZIP proteins upregulated between the PMA-iono, RMD, Bryo and RB stimulation conditions. The only shared

TF upregulated by all LRA stimulations, *BATF3*, also belongs to the bZIP family. The U3 element of the HIV-1 consensus sequences was analyzed for bZIP binding sites (Figure 4.4D). We observed that each promoter had a heterogenous distribution of bZIP half-sites. Heterodimeric sites, which account for the high specificity of bZIP proteins, were surprisingly absent in the HIV-1 promoter.

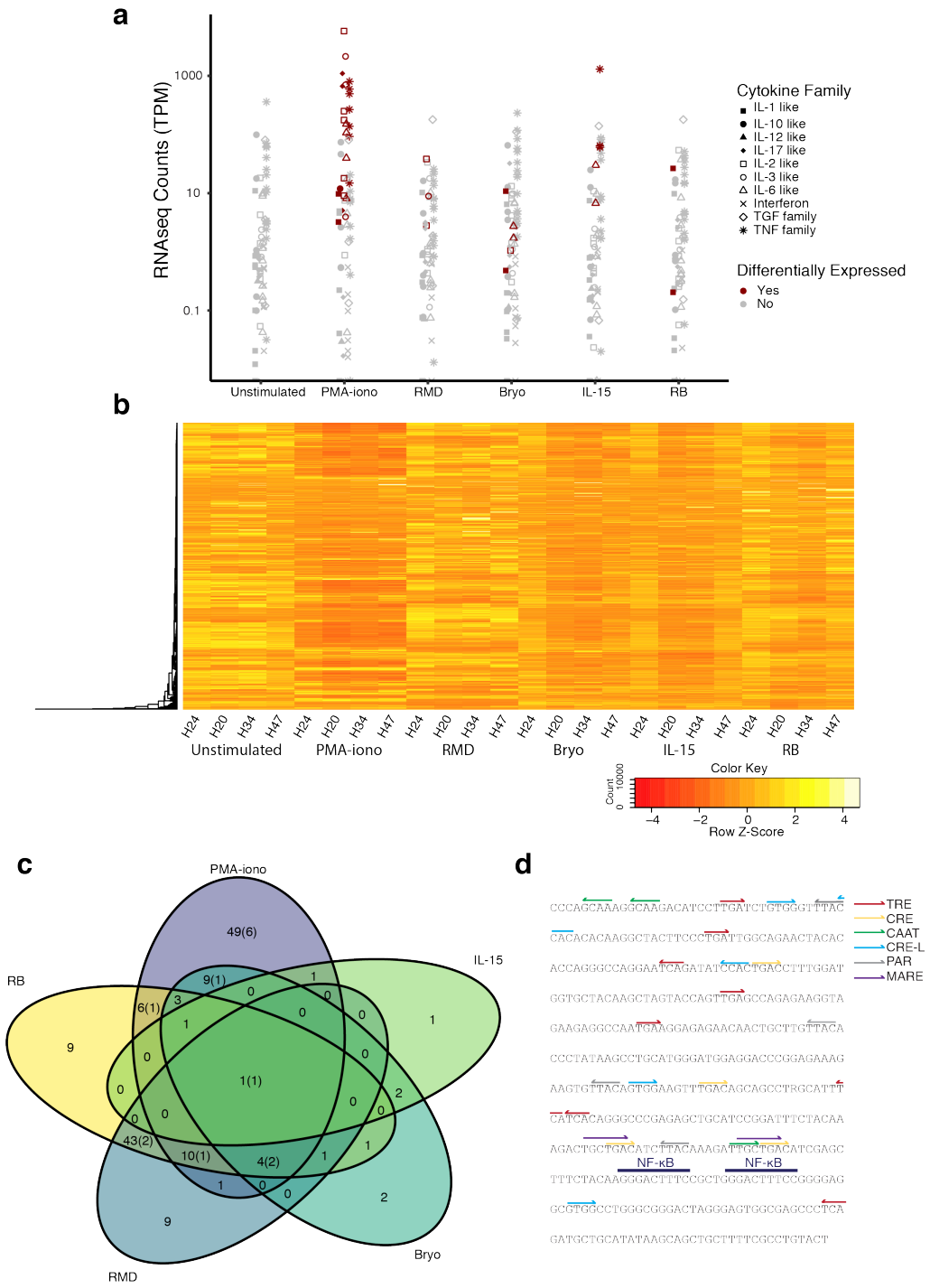

Figure 4.4*: Stimulation with latency reversing agents upregulate different transcription factor profiles. a) RNAseq expression of all human cytokines sorted by family plotted by stimulation condition. RNAseq expression shown in transcripts per million (TPM). Cytokines are color-coded by differentially expressed or not for each LRA stimulation compared to unstimulated based on a significant threshold of 0.01 and a log2 fold-change > 2. b) A*

heatmap constructed of all expression values in TPM of all known human transcription factors. The horizontal dendrogram is described the hierarchical clustering of transcription factors (TF) within these samples. c) The overlap of differentially expressed TF within each of the LRA stimulation conditions is shown by Venn diagram. Differentially expressed TFs were determined by DESeq2 based on a significant threshold of 0.01 and a log2 fold-change > 2. In parenthesis are bZIP genes if applicable. d) Potential binding sites of bZIP genes within the HIV-1 promoter are depicted by half-arrows. NF-κB sites are shown for comparison. The U3 region from H47 consensus sequence was shown as representative sample.

CaptureSeq enables quantification of HIV-1 RNAseq coverage

Overlapping 80-nt tiling probes were designed to enrich the HIV-1 genome and ten human genes for comparison. HIV-1 sequences were derived from single genome amplification from the study participants. The control genes were selected to quantify low to medium expression coverage. Genes that were expressed in each unenriched library were selected to ensure the enrichment would work for each gene. We were able to obtain enriched libraries for each stimulation condition for three of four participants. A representative sample showed an up-shifted linear relationship between the log-transformed expression of target genes including HIV-1 when comparing the unenriched counts to enriched counts, indicating proportional enrichment (Fig. 4.5A). The fraction of target reads was compared to depth of sequencing for the enriched sequencing libraries in order to determine if the sequencing depth biased the enrichment. We observed no statistically significant correlation between sequencing depth of the enriched library and the fraction of target reads for any participant (Fig. 4.5B). After normalization to transcripts per million (TPM) for both the enriched and unenriched libraries, the distribution of fold-enrichments for each gene in each condition was the same for all three participants (Fig. 4.5C). Another potential source of bias was the abundance of the target gene in the unenriched library. When correlating the fold-enrichment and the count of the gene in the unenriched libraries, a

slight correlation was observed (R^2=0.07, p=0.0001). Since this bias contributed to less than 10% of the variance, it was not used to normalize the data any further.

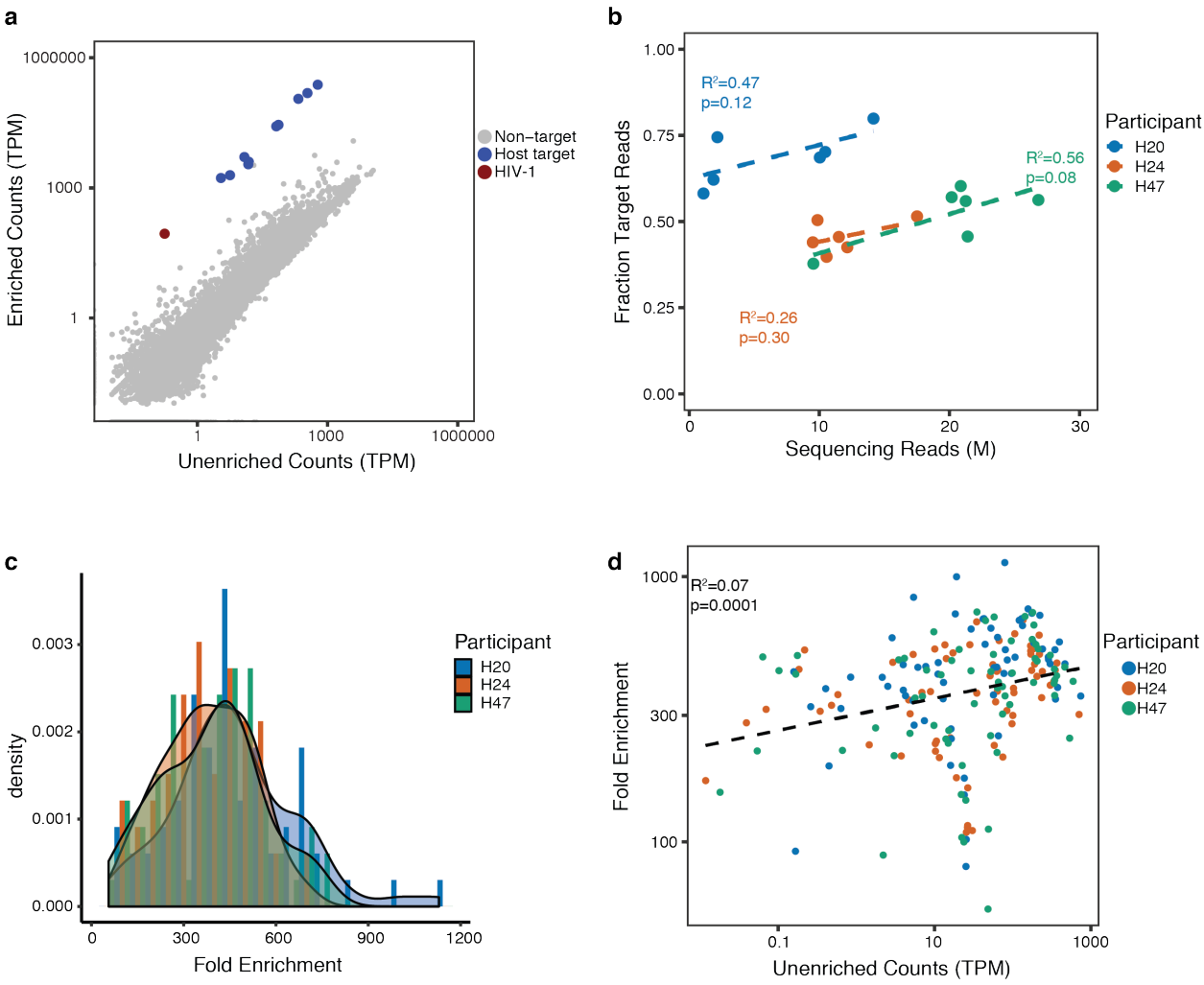

***Figure 4.5**: CaptureSeq proportionally enriches HIV-1 RNAseq reads from latently infected cells. a) Unenriched counts compared to enriched counts for a representative sample (H47-P). Grey dots represent all human genes that were not targeted for enrichment, blue are human genes that were targeted for enrichment by tiling probes and red is HIV-1. b) Fraction of target reads, both host and HIV-1 were plotted against depth of sequencing of the enriched libraries. R^2 is the Pearson's coefficient of determination. c) A histogram with overlaid density plot exhibiting the distribution of fold enrichment for each target gene in each condition for the three participants. d) The fold*

enrichment was plotted against the counts of the target genes in the unenriched libraries. R^2 is the Pearson's coefficient of determination.

After enrichment, reads were observed across the entire HIV-1 genome. Raw reads were plotted along the HIV-1 genome for each condition using the numbering of the consensus sequence for each participant. Due to the primers used to amplify the near-full length sequence, the consensus sequence start was at the end of the U5 element of the 5'UTR. As would be expected by the HIV-1 splicing pattern, the 3'end was more well-covered as it is shared among all HIV-1 transcripts (Fig. 4.6A). We quantified the normalized expression of HIV-1 for each participant. HIV-1 RNA was expressed after stimulation with a variety of latency reversing agents, with RB stimulation having the highest average expression of HIV-1 at 144.25 TPM after enrichment (Fig. 4.6B). Next, we compared HIV-1 caRNA as measured by RT-qPCR to enriched RNAseq HIV-1 coverage. We found distinct relationships in each of the three participants. Only H47 had a statistically significant relationship between the log-transformed RNAseq reads and caRNA (Fig. 4.6C). Differences in the relationship could be due to differences in viral genetics, starting reservoir and intraparticipant differences in response to LRAs. Finally, the enriched RNAseq data was used to interrogate the effect of reactivation on HIV-1 splicing. Although HIV-1 RNAseq reads were not abundant enough to identify splice junctions with high confidence, we could use coverage of different regions of the genome to infer changes in splicing. We calculated the ratio of reads covering the unspliced (US), singly spliced (SS) and multiply spliced (MS) regions of the HIV-1 genome and compared them to caRNA. It is important to note that the coverage of these regions does not exclusively measure the coverage of SS or MS transcripts because HIV-1 has overlapping transcripts. Therefore, the ratio of coverage of these regions must be used to compare changes in splicing. No relationship between any ratio of the splicing

regions was observed with the level of caRNA (Fig. 4.6D). When comparing the ratio of coverage of splice regions and enriched RNAseq counts, the only significant relationship observed was between MS and SS HIV-1 RNA. The inverse relationship indicates there may be competition for the splice machinery between these two groups of transcripts. Overall, we conclude that splicing is not a rate-limiting step of reactivation and more closely resembles zero-order kinetics.

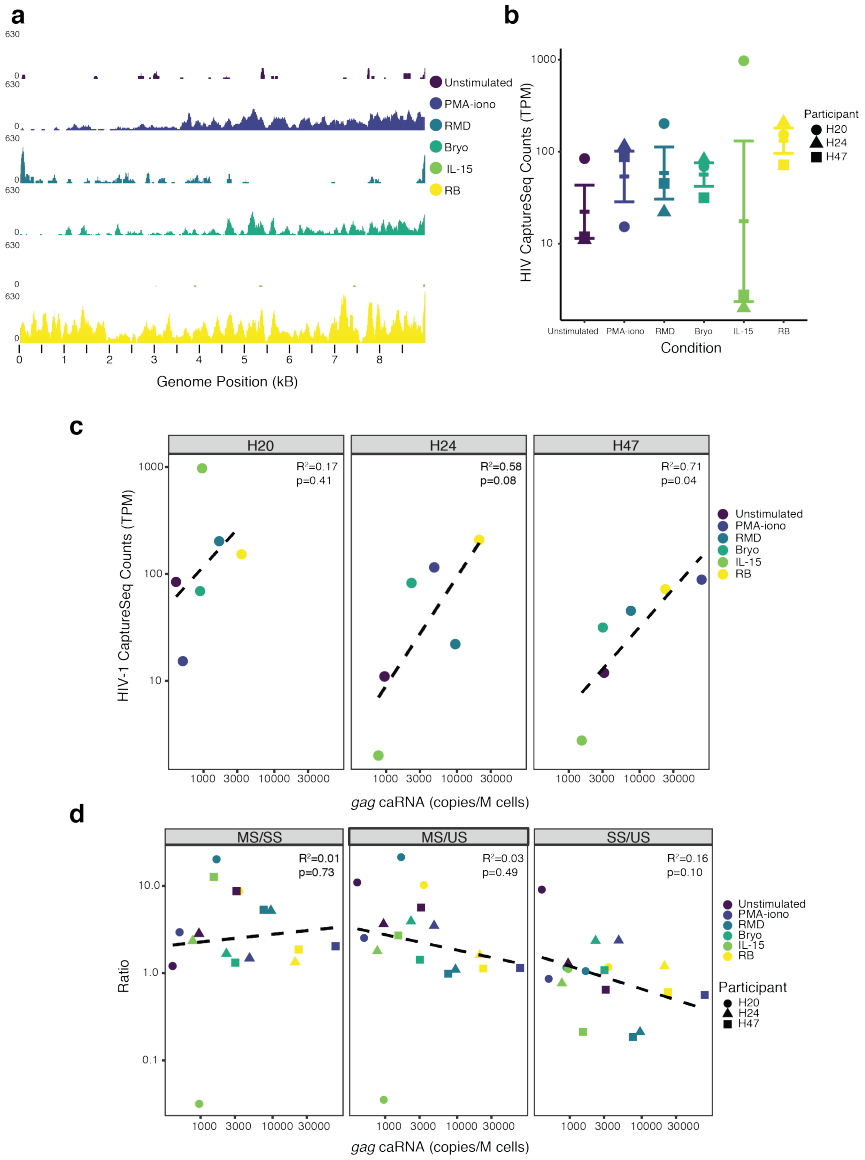

***Figure 4.6**: HIV-1 CaptureSeq reveal heterogeneity in HIV-1 RNA expression and splicing. a) Coverage of HIV-1 consensus sequence for representative sample (H24) plotted by condition. The reads were aligned with Bowtie2 and the unnormalized coverage was visualized with IGV. Numbering of the HIV-1 consensus sequence begins at the end of the U5 element in the 5'UTR. b) The average expression of HIV-1 RNA in TPM was plotted for each condition. Individual data points are shown in addition to mean and error bars denoting s.e. c) Enriched RNAseq counts were compared to gag caRNA copies. Each participant has a separate plot. R^2 is the Pearson's coefficient of determination. d) The ratio of coverage in the unspliced (US), singly spliced (SS) and multiply spliced (MS) regions of the HIV-1 genome was compared to gag caRNA.*

4.4 Discussion

The goal of nn-rCD4 isolation was to increase the probability of CaptureSeq working for latently infected cells. During experimental design, it was not certain whether the enrichment would overcome the rarity of the latently infected cells. Since the initial experimental design and sample collection, it has been found that naïve $CD4^+$ T cells contribute more to the latent reservoir than previously appreciated[87]. Given that enrichment by CaptureSeq is orders of magnitude greater than the enrichment from the nn-r$CD4^+$ T cell isolation, further studies can use this technique to study the latent reservoir in any T cell subset of interest. Bradley *et al.* performed scRNAseq on a latency model and found that viral downregulation was correlated with differentiation of the infected $CD4^+$ T cells[207]. Grau-Expósito et al. found that combinations of LRAs had variable effects on different $CD4^+$ T cell subsets[218]. CaptureSeq could be a useful technique to expand upon these finding and generate hypotheses as to which cellular genes help explain these observations.

Probes were designed with HIV-1 sequences derived from the participants in order to ensure that the probes would hybridize properly to the rare transcripts and enrich. However, the enrichment was more robust than we anticipated, so it is possible that the extra effort to design participant-specific probes is not necessary. If probes designed with a reference sequence lose some enrichment efficiency, it could also be possible to try probes that are clade specific. The host genes that were chosen for enrichment were meant to quantify over a range of values. Genes with medium and low expression that were detected in each unenriched condition were selected. However, the actual values for HIV-1 were lower than any of the control genes. Due to the higher expression of the control genes, there was an enrichment bias towards the genes with

higher expression in the unenriched starting pool. Because the bias contributed less than 10% of the variance, we decided not to normalize. In future experiments, selected host genes with a starting abundance closer to that of the gene or genes of interest is a crucial consideration. Genes that have a better approximation of the target expression could be selected by sequencing an initial experiment very deeply in order to select appropriate genes.

bZIP proteins that bind to the AP-1 site have been shown to enhance HIV-1 transcription[27,219]. There has been little research on the role of the AP-1 binding sites during reactivation from latency. The AP-1 sites described are the half-sites of the CRE binding proteins. However, bZIP proteins must dimerize, as both homodimers and an extensive network of heterodimers, in order to bind to DNA[220]. Binding to half-sites is much lower affinity than full dimeric sites. HIV-1 has many half-sites, which could replace the need for a specific dimer pair with many low affinity sites. The promoter of HIV-1 has commonly been compared to the promoter of cytokines because many of the same factors, namely AP-1, NFAT and NF-κB, are shared[213]. Cytokines, which typically have dimeric bZIP binding sites in their promoters, were specifically upregulated in the PMA-iono stimulation condition but HIV-1 was highly expressed after stimulation with PMA-iono, RMD, Bryo and RB. A difference in the breadth of binding capacity of the whole family of bZIP proteins could explain this difference between the promoter of cytokines and HIV-1.

The dimeric sites though require a specific pair of bZIP proteins, making the dimeric sites more specific than the half-sites. Two studies have shown the impact of these sites on latency. First, if the AP-1 site is extended from a half-site to a dimeric site, latency is promoted and reactivation

is dependent on the JNK pathway[213]. We hypothesize that this is explained by the binding site going from a general specificity to relying on a specific dimer of bZIP proteins. Another study shows that bZIP phosphorylation is required for HIV-1 reactivation, even in the context of nuclear localization of NF-κB[212]. Finally, *BAFT3*, the only gene that was upregulated in each LRA stimulation, was also found to be preferentially upregulated in primary infected cells that successfully reactivated[209]. Based on these studies and the differential expression analysis, bZIP proteins should be investigated further for their role in reactivation.

Several bZIP genes, most prominently Fos and Jun, are regulated post-translationally through phosphorylation as well as through transcriptional upregulation[221]. Therefore, the activity of some of these bZIP genes made be changing after LRA stimulation but would be undetectable by RNAseq. The phosphorylation state of bZIP proteins, particularly in the Jun family, should be studied in future work to determine if LRA stimulation increases their activity.

Yukl *et al*. developed a qPCR-based assay to test the expression of a variety of HIV-1 RNA species. This method was used to test potential steps of post-transcriptional regulation[96]. Interestingly, splicing was one of the regulated steps. It was found in a subsequent study that different LRAs, including HDACi such as romidepsin, showed different post-transcriptional regulation patterns compared to TCR stimulation[222]. We therefore hypothesized that splicing could be a rate-limiting step in reactivation and the splice region coverage of the CaptureSeq data would be dependent on the overall expression of HIV-1 RNA. However, we found no relationship between the any splice ratio and the amount of caRNA measured by RT-qPCR. Our conclusion is that splicing occurs with any amount of HIV-1 RNA, and splice site usage is most

likely set by the virus rather than cellular factors. The key difference between our work and the previously reported studies is that we are measuring the ratio of splicing classes after poly-adenylation. The qPCR-based assay measures abundance of a single splice product in reference to poly-adenylation. The two approaches measure different phenomena and one way to interpret the two findings would be that total splicing might be differentially regulated, but of the ratio of different HIV-1 splice products is not a rate-limiting step as transcription increases.

One of the limitations of this study is that bulk RNAseq was used. Therefore, any differences in response to the LRAs in infected and uninfected cells are lost. Another consideration is that the majority of proviruses are defective[101,223]. Some of these proviruses can be reactivated and transcribe a number of RNA species depending on the nature of the defect in the proviral DNA[108]. This study could not account for defects such as deletions in the 5' LTR that impact transcription and large internal deletions that might impact genome coverage. The assumption of this study is that the defects will be roughly the same per sample in each participant. However, we are not sensitive to LRA specific effects on reactivation of defective proviruses. Likewise, this method is not sensitive to differences in integration site. The integration site might affect how susceptible the provirus is to reactivation[224]. Integration sites have LRA specific effects as well, as integration into a gene that is upregulated or downregulated after stimulation could impact proviral transcription[224]. The final limitation of the current study is a small sample size. Future work will be devoted to expanding on these findings.

In summary, CaptureSeq is capable of achieving proportional enrichment of RNAseq reads across the entire HIV-1 genome. Using the host transcriptomic data, we were able to identify

shared transcription factors in the bZIP family that correlated with HIV-1 expression. Finally, we showed that the ratio of expression between the three main classes of HIV-1 transcripts is not dependent on overall HIV-1 transcription. This technique and these findings can be used to evaluate novel LRAs and more combinations to identify the core determinants of successful latency reversal.

Chapter 5- Discussion

The goal of this dissertation was to develop next-generation sequencing tools to study novel aspects of HIV-1 biology. We used DMS-MaPseq in order to investigate HIV-1 RNA structure and CaptureSeq to examine reactivation of HIV-1 latency. We used DMS-MaPseq in conjunction with a novel clustering algorithm, DREEM, to identify alternate RNA structures in the HIV-1 genome. The development of this strategy has broad implications for the study of RNA biology. RNA structure analysis is a relatively unexplored space in biology due to the limitations of current techniques. One of the biggest obstacles of RNA structure analysis is that RNAs do not have a clearly defined structure. Rather, each RNA species is able to form an ensemble of structures driven by a landscape of thermodynamic minima that are stochastically sampled during transcription. This problem is compounded by the fact that RNA structure is extremely sensitive to changes in the environment, and so folding can be greatly changed depending on the experimental conditions. RNA folding in a cell in the presence of a variety of RNA binding proteins and RNA helicases is different than folding *in vitro*. By resolving alternative RNA structures, DREEM is able to detect more biologically functional RNA structures than other techniques. Combining DREEM with DMS-MaPseq creates a high-throughput assay that can measure intracellular alternative RNA structure. The development of this assay is significant because it greatly reduces the difficulty of identifying alternative RNA structures in a biologically relevant context.

The utility of DMS-MaPseq/DREEM was highlighted by the discovery of a novel splicing regulatory mechanism involving alternative RNA structure at the A3 splice acceptor site. This finding provides evidence for a long-standing hypothesis in the splicing field. Understanding how RNA structure can regulate splicing has therapeutic implications. The barrier for targeting

splicing regulation in HIV-1 is high due to the abundance of potent antiretroviral drugs. However, RNA structures that regulate splicing could be a target for other viruses or diseases for which there are fewer treatment options. Splice regulation could be a prime target if a viral or human protein is involved in the recognition of an RNA structure. In this case, an inhibitor could disrupt the binding of RNA and protein. Identification of the RNA structure and the effects of disrupting the interaction with a protein regulator would be amenable to study using DMS-MaPseq/DREEM.

We developed CaptureSeq for HIV-1 in order to analyze HIV-1 RNA expression in primary, latently infected cells. The latent reservoir is extremely difficult to study due to the rarity of latently infected cells. However, models of latency fail to capture the complex regulation of HIV-1 transcription in latently infected cells. CaptureSeq addresses both of these problems by enriching HIV-1 RNA and RNA from a set of control human genes in order to quantify the differences in HIV-1 expression after treatment with LRAs. This tool can be used to further evaluate LRA candidates and combinations. In contrast to the work on RNA structure, this technique is directly applicable to the development of HIV-1 therapeutics. Due to the ability to simultaneously measure HIV-1 and host response to LRA stimulations, this technique has the promise to both evaluate LRAs and also uncover previously undescribed steps and regulators of HIV-1 latency reversal, such as the bZIP transcription factors identified in the study presented in Chapter 4. This insight is critical as researchers reexamine the biology of HIV-1 latency reversal in order to design LRAs that will be successful in the lab and in the clinic.

DMS-MaPseq/DREEM and CaptureSeq have potential beyond the work presented. In this discussion, I will outline future work to be done to refine these methods, apply them to other research questions and expand the findings of these studies with other experimental methods.

DREEM refinement

One of the biggest limitations of the DMS-MaPseq/DREEM study was stopping clustering at an appropriate number of clusters. Without a clustering limit, the EM algorithm will add a cluster until each read has its own cluster. The Bayesian information criterion (BIC) test is used to determine if adding more clusters is reducing the total error in the system. In the study, the maximum possible number of clusters was set at two, and the BIC test was used only to ensure that the second cluster added information. The exception was the adenine riboswitch (*add*), which was limited only by the BIC. Structures were detected that had not been described in previous literature but contained elements of known structures in different combinations when we allowed for more than two clusters. Even though the BIC provided a conservative cutoff, it is difficult to know when the clustering stops giving biologically relevant structures. In future work, it will be important to assess the significance of structural models that are derived from clustering constrained only by the BIC test.

Overclustering was an important issue when it came to the 5'UTR of HIV-1. Constraining the clustering to just two clusters did not allow the 5'UTR to be analyzed; the first difference that came out was the primer binding site being either completely open or completely closed, which indicated only if the tRNA was bound or not. We expected that there are multiple other conformations, but allowing for more clusters yielded models that did not match when

comparing the whole-genome and the targeted PCR approach. This discrepancy was likely due to the different length of the reads. RRE and A3 matched in both the genome-wide and target PCR approach, probably because there are few conformations and so the difference in length of reads was not such an issue. The reason that more conformations are expected for the 5'UTR is because it is involved in more functions than other RNA structures such as the RRE. Some of the functions could involve structural change, such as dimerization.

Long-read sequencing, such as PacBio and Oxford Nanopore, will become more prevalent as the error-rate of those technologies is reduced. Currently, the error-rate is too high for DMS-MaPseq; the raw error-rate of > 10% for both these technologies is higher than the DMS modification rate of ~2%[225]. DMS-MaPseq could get more information out of longer reads but there are some important considerations. One is the DMS tends to fragment RNA, so again the integrity of the RNA must be balanced with the ability to modify as much as possible to make the most of the reads. Also, structural predictions based on larger windows will be more suited to find large RNA structure changes, however smaller changes will be masked. Being able to interpret as many clusters as possible will be important for DREEM to be able to take advantage of this new sequencing technology.

Alternative RNA structure and splicing regulation beyond the A3 splice acceptor site

In the DMS-MaPseq/DREEM study, alternative structure models for the A4/5 splice sites were made that resembled the structures at A3 splice acceptor site that regulated splicing. A3 was amenable to mutations because a stop codon could be placed in *vpr*, the gene in which A3 is located. A4/5 is located in *tat/rev* genes, which are not dispensable as *vpr* is. However, it is

important to verify that the alternative RNA structures are biologically relevant. One way to get around this pitfall is to insert stop codons in *tat* and *rev*, but supply them to the transfected cell *in trans* on separate plasmids. The splicing assay and DMS-probing could then be done on the mutant strains.

The study by Takata *et al*. that produced HIV-1 viruses with large stretches of synonymous mutations was critical to the forming the hypothesis that alternative RNA structure could regulate splicing. The over-splicing phenotype was observed at A1, A2 and A3 splice acceptor sites, and was most pronounced at A1 and A2[74]. In the genome-wide dataset, A1 and A2 were located in regions of the genome in which low coverage of the reads affected the ability to cluster the data. It would be interesting to sequence with enrichment in order to get coverage in those regions in order to make structural models.

Splice site occlusion at the acceptor sites is not the only way in which RNA structure can interact with splicing. Another hypothesis is that the splice donor sites can also be regulated by alternative RNA structure. In the DMS-MaPseq/DREEM study, we investigated two important splice donor sites, D1 and D4. D1 is the major splice donor site, and needs to be used for all spliced products. We found evidence of heterogeneity at D1 and could propose a similar model, however the 5'UTR had greater heterogeneity then DREEM could handle and so this model will need to be revisited with future iterations of DREEM. In contrast, there was little heterogeneity detected at the D4 splice donor site, which is used in all multiply spliced HIV-1 transcripts. The D4 splice site was always exposed. This result indicates that the regulation is governed either entirely by enhancers and silencers, or that RNA structure at the A7 splice acceptor site is a

regulatory element. A7 seemed to have the occluded/exposed alternative structures however longer reads were needed to make better models at that site since the silencer and enhancer sites were further away from the splice site than A3. This splice site would be interesting to study because it is different than the other slice donor/acceptor pairs in the genome.

In addition to disrupting regulation of known splice sites, Tanaka *et al*. uncovered cryptic splice sites in the Gag-Pol coding region with the use of synonymous mutants. This experiment suggested that RNA structure is used to occlude the cryptic splice sites under normal conditions. We can utilize DMS-MaPseq and DREEM in order to uncover the RNA structure that is able to ensure that these cryptic splice sites remain occluded.

RNA structure is susceptible to changes in temperature, as higher temperature can break bonds and reduce thermodynamic stability of RNA conformations. Interestingly, HIV-1 splicing was misregulated when infected cells were incubated at 42^{o}C[64]. The virus over-spliced at A1 and A2. We hypothesize inhibitory RNA structures are unable to form at the higher temperature. We could utilize DMS-MaPseq and DREEM to test this hypothesis.

Implications of alternative RNA structure on splicing of human RNA

Human genes also undergo alternative splicing, although the exons are not overlapping as they are in the HIV-1 genome. The current understanding of splicing regulation within the human genome is similar to what was known about HIV-1 before our work. Each splice donor and acceptor site have splice enhance and silencer elements that control their usage. Families of RNA

binding proteins recognize conserved binding sequences in order to prevent or recruit the binding of splice machinery. These elements can be located in the intron or exon.

Although the human genome is not under the same length constraint as the HIV-1 genome, it is possible that human genes also use alternative RNA structure to regulate splicing in addition to enhancer and silencer elements. A few examples of RNA structure impacting human splicing exist, although these examples are still not well-understood. One of the better understood examples involves the disease spinal muscular atrophy, which can be treated with RNA therapeutics that alter splice usage by preventing RNA structure from forming at a splice site. In people with spinal muscular atrophy, the gene survival of motor neuron 1 (*SMN1*) is defective due to mutation or internal deletions[226]. *SMN2*, a closely related gene that is a product of gene duplication, has a single nucleotide polymorphism in exon 7 that prevents inclusion of exon 7 in the transcript. The antisense RNA therapeutic prevents a long-distance stem loop from occurring at the exon 7 splice acceptor site in *SMN2*, which promotes inclusion of exon 7[226]. *SMN2* with exon 7 included can partially restore the lost function of *SMN1*, providing substantive clinical benefit. In this example, RNA structure regulates the skipping of the exon however there is no known role for alternative RNA structure. However, recent research has shown the splicing pattern of *SMN2* changes under conditions of oxidative stress, including decreased exon 7 inclusion[227]. The inability to break the RNA structure by ATP-dependent RNA helicases during oxidative stress has been proposed to explain the condition-specific changes to splice regulation, but it could also be that the thermodynamics of RNA structure are responsible for this observation. This example shows that RNA structure can be used by cells to regulate human splicing.

Our hypothesis is that RNA viruses such as HIV-1 represent extreme examples of alternative RNA structure usage for splicing regulation, but human genes probably use this mechanism more widely than currently known. The development of DMS-MaPseq/DREEM provides a powerful tool to interrogate the relationship between RNA structure and splicing.

HIV-1 CaptureSeq without poly-adenylation selection

One of the main conclusions from the CaptureSeq study was that splice usage is not dependent on expression level of HIV-1 RNA. This finding is a step beyond the finding of an assay that compares levels of different RT-qPCR products across the HIV-1 genome to infer regulation of steps from transcription initiation and elongation through to poly-adenylation and splicing[96]. The selection of poly-adenylated transcripts in the CaptureSeq library generation is what sets these two studies apart. However, if the CaptureSeq were redone without selection for poly-adenylated transcripts, a more direct comparison could be made. Instead of poly-adenylation selection, rRNA subtraction could be performed during library generation. Genome coverage of the enriched HIV-1 RNA would provide insight into transcription initiation and elongation by comparing the peak size of the 5'UTR to coverage in the *gag-pol* region. Comparing the poly-adenylation CaptureSeq coverage to the rRNA subtracted coverage would be a way to explore the regulated steps of HIV-1 transcription while also comparing host transcriptomic information.

HIV-1 and AP-1 binding

The HIV-1 promoter is a complex regulatory element. AP-1 sites, which are bound by dimers of bZIP proteins such as Jun and Fos, have been shown to enhance HIV-1 transcription[27,219]. bZIP

proteins have been shown to have a role in latency and reactivation, although it is unclear if bZIP proteins are necessary for reactivation or just enhance reactivation. Based on the analysis of host transcriptomic data from the CaptureSeq dataset, it was found that many bZIP genes were upregulated and that the only shared upregulated transcription factor between all LRA stimulations was the bZIP gene *BATF3*. The HIV-1 promoters of the three HEAL participants were mapped for half-sites of bZIP genes, and more potential binding sites were identified than previously recognized. However, the half-sites are likely low affinity sites, and so we hypothesize that the HIV-1 promoter makes up for the low affinity site by having many sites. Therefore, many different bZIP dimers can bind to the promoter depending on which of these genes are upregulated in the cell at the time.

To test this hypothesis more directly, a novel technology called Caspex could be utilized in an HIV-1 latency model[228]. For this assay, a T cell line such as Jurkat is stably transfected with a plasmid containing a catalytically-dead Cas protein fused to a labeling domain by a flexible linker along with a guide RNA designed against the HIV-1 promoter. The cells are then infected with a virus with two fluorescent tags, one under control of the HIV-1 promoter and one with a constitutively active promoter[229]. The cells that express the constitutively active tag but not the HIV-1 promoter-dependent tag are selected and used for stimulation with LRAs. During stimulation, the guide RNA brings the Capsex complex to the HIV-1 promoter and any transcription factors binding to the promoter are labeled. The cells are then lysed, labeled proteins are immunoprecipitated and identified by mass spectrometry[228]. This experiment could verify that a diverse set of bZIP dimers are binding to the reactivated HIV-1 promoter and identify shared factors after different LRA stimulations. This experiment would also address one

of the key limitations of the RNAseq study, which is identifying important transcription factors for reactivation that are post-translationally regulated.

HIV-1 RNA structure and latency

Now that DMS-MaPseq has been adapted for productively HIV-1 infected cells, it could be used to study latently infected cells as well. One hypothesis that could be tested with DMS-MaPseq and DREEM is that the 5'UTR of HIV-1 could be inhibitory to reactivation of the virus. If the 5'UTR is inhibiting translation, then the RNA cannot be translated into HIV-1 proteins such as Tat and Rev that sustain infection of the cell. If translation of HIV-1 transcripts is minimal, only low amounts of HIV-1 antigen are available to be presented to cytotoxic lymphocytes. The goal of 'shock and kill' is to target transcriptional repression of the HIV-1 provirus. If there is an essential RNA helicase that helps to unwind the inhibitory 5'UTR structure that is not upregulated, then an increase in HIV-1 transcription may not be sufficient for reactivation. HIV-1 RNA could be transcribed, but without the Tat positive feedback loop, the provirus could become chromatinized again before productive infection begins. A two-structure equilibrium has been proposed for the HIV-1 5'UTR, one which promotes translation and one that inhibits translation, allowing for packaging[41,45]. DMS-MaPseq and DREEM would be able to detect the relative proportion of each form in latently infected cells treated with different LRAs. This experiment would identify if the RNA structure of the 5'UTR is a factor in the successful reactivation of HIV-1 from latency. If the proportion of 5'UTR structures does change with the different LRAs, the next step would be to identify the essential RNA helicases that are upregulated by T-cell activation that contribute to successful reactivation.

In conclusion, we developed two next-generation sequencing techniques, DMS-MaPseq and CaptureSeq, to answer pressing questions about HIV-1. DMS-MaPseq together with the alternative RNA structure detection algorithm DREEM represent the cutting edge of RNA structure analysis. We used these tools to identify a novel HIV-1 RNA structure and determine the extent of RNA structure heterogeneity across the HIV-1 genome. The findings have impacts on both the study of HIV-1 and on RNA biology more broadly. CaptureSeq was adapted to study HIV-1 latency, which overcomes a major obstacle in using RNAseq to study HIV-1 gene expression. We showed that CaptureSeq is a valuable tool for determining how HIV-1 reactivation occurs in latently infected cells after stimulation with a diverse array of LRAs.

References

1 UNAIDS. 2018 Global HIV Statistics. (2019).

2 Wilen, C. B., Tilton, J. C. & Doms, R. W. HIV: cell binding and entry. *Cold Spring Harb Perspect Med* **2**, doi:10.1101/cshperspect.a006866 (2012).

3 Ratner, L. *et al.* Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature* **313**, 277-284, doi:10.1038/313277a0 (1985).

4 Telesnitsky, A. & Goff, S. P. in *Retroviruses* (eds J. M. Coffin, S. H. Hughes, & H. E. Varmus) (1997).

5 Gotte, M., Li, X. & Wainberg, M. A. HIV-1 reverse transcription: a brief overview focused on structure-function relationships among molecules involved in initiation of the reaction. *Arch Biochem Biophys* **365**, 199-210, doi:10.1006/abbi.1999.1209 (1999).

6 Hu, W. S. & Hughes, S. H. HIV-1 reverse transcription. *Cold Spring Harb Perspect Med* **2**, doi:10.1101/cshperspect.a006882 (2012).

7 Abram, M. E. *et al.* Mutations in HIV-1 reverse transcriptase affect the errors made in a single cycle of viral replication. *J Virol* **88**, 7589-7601, doi:10.1128/JVI.00302-14 (2014).

8 Abram, M. E., Ferris, A. L., Shao, W., Alvord, W. G. & Hughes, S. H. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *J Virol* **84**, 9864-9878, doi:10.1128/JVI.00915-10 (2010).

9 Cuevas, J. M., Geller, R., Garijo, R., Lopez-Aldeguer, J. & Sanjuan, R. Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS Biol* **13**, e1002251, doi:10.1371/journal.pbio.1002251 (2015).

10 Sarafianos, S. G. *et al.* Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition. *J Mol Biol* **385**, 693-713, doi:10.1016/j.jmb.2008.10.071 (2009).

11 Lee, K. *et al.* Flexible use of nuclear import pathways by HIV-1. *Cell Host Microbe* **7**, 221-233, doi:10.1016/j.chom.2010.02.007 (2010).

12 Schaller, T. *et al.* HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. *PLoS Pathog* **7**, e1002439, doi:10.1371/journal.ppat.1002439 (2011).

13 De Iaco, A. *et al.* TNPO3 protects HIV-1 replication from CPSF6-mediated capsid stabilization in the host cell cytoplasm. *Retrovirology* **10**, 20, doi:10.1186/1742-4690-10-20 (2013).

14 Sowd, G. A. *et al.* A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc Natl Acad Sci U S A* **113**, E1054-1063, doi:10.1073/pnas.1524213113 (2016).

15 Cherepanov, P. *et al.* HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J Biol Chem* **278**, 372-381, doi:10.1074/jbc.M209278200 (2003).

16 Maertens, G. *et al.* LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. *J Biol Chem* **278**, 33528-33539, doi:10.1074/jbc.M303594200 (2003).

17 Ciuffi, A. *et al.* A role for LEDGF/p75 in targeting HIV DNA integration. *Nat Med* **11**, 1287-1289, doi:10.1038/nm1329 (2005).

18 Llano, M. *et al.* An essential role for LEDGF/p75 in HIV integration. *Science* **314**, 461-464, doi:10.1126/science.1132319 (2006).

19 Shun, M. C. *et al.* LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes Dev* **21**, 1767-1778, doi:10.1101/gad.1565107 (2007).

20 Engelman, A., Mizuuchi, K. & Craigie, R. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. *Cell* **67**, 1211-1221, doi:10.1016/0092-8674(91)90297-c (1991).

21 Arya, S. K., Guo, C., Josephs, S. F. & Wong-Staal, F. Trans-activator gene of human T-lymphotropic virus type III (HTLV-III). *Science* **229**, 69-73, doi:10.1126/science.2990040 (1985).

22 Sodroski, J., Patarca, R., Rosen, C., Wong-Staal, F. & Haseltine, W. Location of the trans-activating region on the genome of human T-cell lymphotropic virus type III. *Science* **229**, 74-77, doi:10.1126/science.2990041 (1985).

23 Selby, M. J., Bain, E. S., Luciw, P. A. & Peterlin, B. M. Structure, sequence, and position of the stem-loop in tar determine transcriptional elongation by tat through the HIV-1 long terminal repeat. *Genes Dev* **3**, 547-558, doi:10.1101/gad.3.4.547 (1989).

24 Herrmann, C. H. & Rice, A. P. Specific interaction of the human immunodeficiency virus Tat proteins with a cellular protein kinase. *Virology* **197**, 601-608, doi:10.1006/viro.1993.1634 (1993).

25 Zhu, Y. *et al.* Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation in vitro. *Genes Dev* **11**, 2622-2632, doi:10.1101/gad.11.20.2622 (1997).

26 Perkins, N. D. *et al.* A cooperative interaction between NF-kappa B and Sp1 is required for HIV-1 enhancer activation. *EMBO J* **12**, 3551-3558 (1993).

27 Yang, X., Chen, Y. & Gabuzda, D. ERK MAP kinase links cytokine signals to activation of latent HIV-1 infection by stimulating a cooperative interaction of AP-1 and NF-kappaB. *J Biol Chem* **274**, 27981-27988, doi:10.1074/jbc.274.39.27981 (1999).

28 Nabel, G. & Baltimore, D. An inducible transcription factor activates expression of human immunodeficiency virus in T cells. *Nature* **326**, 711-713, doi:10.1038/326711a0 (1987).

29 Giffin, M. J. *et al.* Structure of NFAT1 bound as a dimer to the HIV-1 LTR kappa B element. *Nat Struct Biol* **10**, 800-806, doi:10.1038/nsb981 (2003).

30 Purcell, D. F. & Martin, M. A. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *J Virol* **67**, 6365-6378 (1993).

31 Ocwieja, K. E. *et al.* Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic Acids Res* **40**, 10345-10355, doi:10.1093/nar/gks753 (2012).

32 Jacks, T. *et al.* Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* **331**, 280-283, doi:10.1038/331280a0 (1988).

33 Wilson, W. *et al.* HIV expression strategies: ribosomal frameshifting is directed by a short sequence in both mammalian and yeast systems. *Cell* **55**, 1159-1169, doi:10.1016/0092-8674(88)90260-7 (1988).

34 Sundquist, W. I. & Krausslich, H. G. HIV-1 assembly, budding, and maturation. *Cold Spring Harb Perspect Med* **2**, a006924, doi:10.1101/cshperspect.a006924 (2012).

35 Barnwal, R. P., Yang, F. & Varani, G. Applications of NMR to structure determination of RNAs large and small. *Arch Biochem Biophys* **628**, 42-56, doi:10.1016/j.abb.2017.06.003 (2017).

36 Abbink, T. E. & Berkhout, B. A novel long distance base-pairing interaction in human immunodeficiency virus type 1 RNA occludes the Gag start codon. *J Biol Chem* **278**, 11601-11611, doi:10.1074/jbc.M210291200 (2003).

37 Berkhout, B. Structure and function of the human immunodeficiency virus leader RNA. *Prog Nucleic Acid Res Mol Biol* **54**, 1-34, doi:10.1016/s0079-6603(08)60359-1 (1996).

38 Heng, X. *et al.* Identification of a minimal region of the HIV-1 5'-leader required for RNA dimerization, NC binding, and packaging. *J Mol Biol* **417**, 224-239, doi:10.1016/j.jmb.2012.01.033 (2012).

39 van Bel, N., Das, A. T., Cornelissen, M., Abbink, T. E. & Berkhout, B. A short sequence motif in the 5' leader of the HIV-1 genome modulates extended RNA dimer formation and virus replication. *J Biol Chem* **289**, 35061-35074, doi:10.1074/jbc.M114.621425 (2014).

40 Russell, R. S., Liang, C. & Wainberg, M. A. Is HIV-1 RNA dimerization a prerequisite for packaging? Yes, no, probably? *Retrovirology* **1**, 23, doi:10.1186/1742-4690-1-23 (2004).

41 Huthoff, H. & Berkhout, B. Two alternating structures of the HIV-1 leader RNA. *RNA* **7**, 143-157, doi:10.1017/s1355838201001881 (2001).

42 Kenyon, J. C., Prestwood, L. J., Le Grice, S. F. & Lever, A. M. In-gel probing of individual RNA conformers within a mixed population reveals a dimerization structural switch in the HIV-1 leader. *Nucleic Acids Res* **41**, e174, doi:10.1093/nar/gkt690 (2013).

43 Lu, K. *et al.* NMR detection of structures in the HIV-1 5'-leader RNA that regulate genome packaging. *Science* **334**, 242-245, doi:10.1126/science.1210460 (2011).

44 Keane, S. C. *et al.* RNA structure. Structure of the HIV-1 RNA packaging signal. *Science* **348**, 917-921, doi:10.1126/science.aaa9266 (2015).

45 Kharytonchyk, S. *et al.* Transcriptional start site heterogeneity modulates the structure and function of the HIV-1 genome. *Proc Natl Acad Sci U S A* **113**, 13378-13383, doi:10.1073/pnas.1616627113 (2016).

46 Brigham, B. S., Kitzrow, J. P., Reyes, J. C., Musier-Forsyth, K. & Munro, J. B. Intrinsic conformational dynamics of the HIV-1 genomic RNA 5'UTR. *Proc Natl Acad Sci U S A* **116**, 10372-10381, doi:10.1073/pnas.1902271116 (2019).

47 Boeras, I. *et al.* The basal translation rate of authentic HIV-1 RNA is regulated by 5'UTR nt-pairings at junction of R and U5. *Sci Rep* **7**, 6902, doi:10.1038/s41598-017-06883-9 (2017).

48 Dethoff, E. A., Petzold, K., Chugh, J., Casiano-Negroni, A. & Al-Hashimi, H. M. Visualizing transient low-populated structures of RNA. *Nature* **491**, 724-728, doi:10.1038/nature11498 (2012).

49 Hadzopoulou-Cladaras, M. *et al.* The rev (trs/art) protein of human immunodeficiency virus type 1 affects viral mRNA and protein expression via a cis-acting sequence in the env region. *J Virol* **63**, 1265-1274 (1989).

50 Hammarskjold, M. L. *et al.* Regulation of human immunodeficiency virus env expression by the rev gene product. *J Virol* **63**, 1959-1966 (1989).

51 Malim, M. H., Hauber, J., Le, S. Y., Maizel, J. V. & Cullen, B. R. The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature* **338**, 254-257, doi:10.1038/338254a0 (1989).

52 Sherpa, C., Rausch, J. W., Le Grice, S. F., Hammarskjold, M. L. & Rekosh, D. The HIV-1 Rev response element (RRE) adopts alternative conformations that promote different rates of virus replication. *Nucleic Acids Res* **43**, 4676-4686, doi:10.1093/nar/gkv313 (2015).

53 Bai, Y., Tambe, A., Zhou, K. & Doudna, J. A. RNA-guided assembly of Rev-RRE nuclear export complexes. *Elife* **3**, e03656, doi:10.7554/eLife.03656 (2014).

54 Weinreb, C. *et al.* 3D RNA and Functional Interactions from Evolutionary Couplings. *Cell* **165**, 963-975, doi:10.1016/j.cell.2016.03.030 (2016).

55 Jackson, P. E., Tebit, D. M., Rekosh, D. & Hammarskjold, M. L. Rev-RRE Functional Activity Differs Substantially Among Primary HIV-1 Isolates. *AIDS Res Hum Retroviruses* **32**, 923-934, doi:10.1089/AID.2016.0047 (2016).

56 Sherpa, C. *et al.* Evolution of the HIV-1 Rev Response Element during Natural Infection Reveals Nucleotide Changes That Correlate with Altered Structure and Increased Activity over Time. *J Virol* **93**, doi:10.1128/JVI.02102-18 (2019).

57 Low, J. T. *et al.* Structure and dynamics of the HIV-1 frameshift element RNA. *Biochemistry* **53**, 4282-4291, doi:10.1021/bi5004926 (2014).

58 Huang, X., Yang, Y., Wang, G., Cheng, Q. & Du, Z. Highly conserved RNA pseudoknots at the Gag-Pol junction of HIV-1 suggest a novel mechanism of -1 ribosomal frameshifting. *RNA* **20**, 587-593, doi:10.1261/rna.042457.113 (2014).

59 Watts, J. M. *et al.* Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711-716, doi:10.1038/nature08237 (2009).

60 Dulude, D., Berchiche, Y. A., Gendron, K., Brakier-Gingras, L. & Heveker, N. Decreasing the frameshift efficiency translates into an equivalent reduction of the replication of the human immunodeficiency virus type 1. *Virology* **345**, 127-136, doi:10.1016/j.virol.2005.08.048 (2006).

61 Garcia-Miranda, P. *et al.* Stability of HIV Frameshift Site RNA Correlates with Frameshift Efficiency and Decreased Virus Infectivity. *J Virol* **90**, 6906-6917, doi:10.1128/JVI.00149-16 (2016).

62 Houck-Loomis, B. *et al.* An equilibrium-dependent retroviral mRNA switch regulates translational recoding. *Nature* **480**, 561-564, doi:10.1038/nature10657 (2011).

63 Liu, Y. *et al.* The roles of five conserved lentiviral RNA structures in HIV-1 replication. *Virology* **514**, 1-8, doi:10.1016/j.virol.2017.10.020 (2018).

64 Emery, A., Zhou, S., Pollom, E. & Swanstrom, R. Characterizing HIV-1 Splicing by Using Next-Generation Sequencing. *J Virol* **91**, doi:10.1128/JVI.02515-16 (2017).

65 Stoltzfus, C. M. & Madsen, J. M. Role of viral splicing elements and cellular RNA binding proteins in regulation of HIV-1 alternative RNA splicing. *Curr HIV Res* **4**, 43-55, doi:10.2174/157016206775197655 (2006).

66 Stoltzfus, C. M. Chapter 1. Regulation of HIV-1 alternative RNA splicing and its role in virus replication. *Adv Virus Res* **74**, 1-40, doi:10.1016/S0065-3527(09)74001-1 (2009).

67 Madsen, J. M. & Stoltzfus, C. M. 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, doi:10.1186/1742-4690-3-10 (2006).

68 Exline, C. M., Feng, Z. & Stoltzfus, C. M. Negative and positive mRNA splicing elements act competitively to regulate human immunodeficiency virus type 1 vif gene expression. *J Virol* **82**, 3921-3931, doi:10.1128/JVI.01558-07 (2008).

69 Kutluay, S. B. *et al.* Genome-Wide Analysis of Heterogeneous Nuclear Ribonucleoprotein (hnRNP) Binding to HIV-1 RNA Reveals a Key Role for hnRNP H1 in Alternative Viral mRNA Splicing. *J Virol* **93**, doi:10.1128/JVI.01048-19 (2019).

70 Kammler, S. *et al.* The strength of the HIV-1 3' splice sites affects Rev function. *Retrovirology* **3**, 89, doi:10.1186/1742-4690-3-89 (2006).

71 Kammler, S. *et al.* 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, doi:10.1017/s1355838201001212 (2001).

72 Chang, D. D. & Sharp, P. A. Regulation by HIV Rev depends upon recognition of splice sites. *Cell* **59**, 789-795, doi:10.1016/0092-8674(89)90602-8 (1989).

73 Stutz, F. & Rosbash, M. A functional interaction between Rev and yeast pre-mRNA is related to splicing complex formation. *EMBO J* **13**, 4096-4104 (1994).

74 Takata, M. A. *et al.* Global synonymous mutagenesis identifies cis-acting RNA elements that regulate HIV-1 splicing and replication. *PLoS Pathog* **14**, e1006824, doi:10.1371/journal.ppat.1006824 (2018).

75 Abbink, T. E. & Berkhout, B. RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor. *J Virol* **82**, 3090-3098, doi:10.1128/JVI.01479-07 (2008).

76 Rollins, C., Levengood, J. D., Rife, B. D., Salemi, M. & Tolbert, B. S. Thermodynamic and phylogenetic insights into hnRNP A1 recognition of the HIV-1 exon splicing silencer 3 element. *Biochemistry* **53**, 2172-2184, doi:10.1021/bi500180p (2014).

77 Siliciano, J. D. *et al.* Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat Med* **9**, 727-728, doi:10.1038/nm880 (2003).

78 Crooks, A. M. *et al.* Precise Quantitation of the Latent HIV-1 Reservoir: Implications for Eradication Strategies. *J Infect Dis* **212**, 1361-1365, doi:10.1093/infdis/jiv218 (2015).

79 Finzi, D. *et al.* Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat Med* **5**, 512-517, doi:10.1038/8394 (1999).

80 Li, J. Z. *et al.* The size of the expressed HIV reservoir predicts timing of viral rebound after treatment interruption. *AIDS* **30**, 343-353, doi:10.1097/QAD.0000000000000953 (2016).

81 Whitney, J. B. *et al.* Rapid seeding of the viral reservoir prior to SIV viraemia in rhesus monkeys. *Nature* **512**, 74-77, doi:10.1038/nature13594 (2014).

82 Colby, D. J. *et al.* Rapid HIV RNA rebound after antiretroviral treatment interruption in persons durably suppressed in Fiebig I acute HIV infection. *Nat Med* **24**, 923-926, doi:10.1038/s41591-018-0026-6 (2018).

83 Chomont, N. *et al.* HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat Med* **15**, 893-900, doi:10.1038/nm.1972 (2009).

84 Mahnke, Y. D., Brodie, T. M., Sallusto, F., Roederer, M. & Lugli, E. The who's who of T-cell differentiation: human memory T-cell subsets. *Eur J Immunol* **43**, 2797-2809, doi:10.1002/eji.201343751 (2013).

85 Soriano-Sarabia, N. *et al.* Quantitation of replication-competent HIV-1 in populations of resting CD4+ T cells. *J Virol* **88**, 14070-14077, doi:10.1128/JVI.01900-14 (2014).

86 Buzon, M. J. *et al.* HIV-1 persistence in CD4+ T cells with stem cell-like properties. *Nat Med* **20**, 139-142, doi:10.1038/nm.3445 (2014).

87 Zerbato, J. M., McMahon, D. K., Sobolewski, M. D., Mellors, J. W. & Sluis-Cremer, N. Naive CD4+ T Cells Harbor a Large Inducible Reservoir of Latent, Replication-competent Human Immunodeficiency Virus Type 1. *Clin Infect Dis* **69**, 1919-1925, doi:10.1093/cid/ciz108 (2019).

88 Verdin, E., Paras, P., Jr. & Van Lint, C. Chromatin disruption in the promoter of human immunodeficiency virus type 1 during transcriptional activation. *EMBO J* **12**, 3249-3259 (1993).

89 Van Lint, C., Emiliani, S. & Verdin, E. The expression of a small fraction of cellular genes is changed in response to histone hyperacetylation. *Gene Expr* **5**, 245-253 (1996).

90 Friedman, J. *et al.* Epigenetic silencing of HIV-1 by the histone H3 lysine 27 methyltransferase enhancer of Zeste 2. *J Virol* **85**, 9078-9089, doi:10.1128/JVI.00836-11 (2011).

91 Nguyen, K., Das, B., Dobrowolski, C. & Karn, J. Multiple Histone Lysine Methyltransferases Are Required for the Establishment and Maintenance of HIV-1 Latency. *MBio* **8**, doi:10.1128/mBio.00133-17 (2017).

92 He, G. & Margolis, D. M. Counterregulation of chromatin deacetylation and histone deacetylase occupancy at the integrated promoter of human immunodeficiency virus type 1 (HIV-1) by the HIV-1 repressor YY1 and HIV-1 activator Tat. *Mol Cell Biol* **22**, 2965-2973, doi:10.1128/mcb.22.9.2965-2973.2002 (2002).

93 Duverger, A. *et al.* Determinants of the establishment of human immunodeficiency virus type 1 latency. *J Virol* **83**, 3078-3093, doi:10.1128/JVI.02058-08 (2009).

94 Chiang, K., Sung, T. L. & Rice, A. P. Regulation of cyclin T1 and HIV-1 Replication by microRNAs in resting CD4+ T lymphocytes. *J Virol* **86**, 3244-3252, doi:10.1128/JVI.05065-11 (2012).

95 Gagne, M. *et al.* Strength of T cell signaling regulates HIV-1 replication and establishment of latency. *PLoS Pathog* **15**, e1007802, doi:10.1371/journal.ppat.1007802 (2019).

96 Yukl, S. A. *et al.* HIV latency in isolated patient CD4(+) T cells may be due to blocks in HIV transcriptional elongation, completion, and splicing. *Sci Transl Med* **10**, doi:10.1126/scitranslmed.aap9927 (2018).

97 Lassen, K. G., Ramyar, K. X., Bailey, J. R., Zhou, Y. & Siliciano, R. F. Nuclear retention of multiply spliced HIV-1 RNA in resting CD4+ T cells. *PLoS Pathog* **2**, e68, doi:10.1371/journal.ppat.0020068 (2006).

98 Wagner, T. A. *et al.* HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345**, 570-573, doi:10.1126/science.1256304 (2014).

99 Maldarelli, F. *et al.* HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345**, 179-183, doi:10.1126/science.1254194 (2014).

100 Wang, Z. *et al.* Expanded cellular clones carrying replication-competent HIV-1 persist, wax, and wane. *Proc Natl Acad Sci U S A* **115**, E2575-E2584, doi:10.1073/pnas.1720665115 (2018).

101 Bruner, K. M. *et al.* Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nat Med* **22**, 1043-1049, doi:10.1038/nm.4156 (2016).

102 Pinzone, M. R. *et al.* Longitudinal HIV sequencing reveals reservoir expression leading to decay which is obscured by clonal expansion. *Nat Commun* **10**, 728, doi:10.1038/s41467-019-08431-7 (2019).

103 Einkauf, K. B. *et al.* Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *J Clin Invest* **129**, 988-998, doi:10.1172/JCI124291 (2019).

104 Coffin, J. M. *et al.* Clones of infected cells arise early in HIV-infected individuals. *JCI Insight* **4**, doi:10.1172/jci.insight.128432 (2019).

105 Bui, J. K. *et al.* Proviruses with identical sequences comprise a large fraction of the replication-competent HIV reservoir. *PLoS Pathog* **13**, e1006283, doi:10.1371/journal.ppat.1006283 (2017).

106 Sheehy, A. M., Gaddis, N. C., Choi, J. D. & Malim, M. H. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418**, 646-650, doi:10.1038/nature00939 (2002).

107 Simon, V. *et al.* Natural variation in Vif: differential impact on APOBEC3G/3F and a potential role in HIV-1 diversification. *PLoS Pathog* **1**, e6, doi:10.1371/journal.ppat.0010006 (2005).

108 Pollack, R. A. *et al.* Defective HIV-1 Proviruses Are Expressed and Can Be Recognized by Cytotoxic T Lymphocytes, which Shape the Proviral Landscape. *Cell Host Microbe* **21**, 494-506 e494, doi:10.1016/j.chom.2017.03.008 (2017).

109 Lehrman, G. *et al.* Depletion of latent HIV-1 infection in vivo: a proof-of-concept study. *Lancet* **366**, 549-555, doi:10.1016/S0140-6736(05)67098-5 (2005).

110 Archin, N. M. *et al.* Administration of vorinostat disrupts HIV-1 latency in patients on antiretroviral therapy. *Nature* **487**, 482-485, doi:10.1038/nature11286 (2012).

111 Deeks, S. G. HIV: Shock and kill. *Nature* **487**, 439-440, doi:10.1038/487439a (2012).

112 Mehla, R. *et al.* Bryostatin modulates latent HIV-1 infection via PKC and AMPK signaling but inhibits acute infection in a receptor independent manner. *PLoS One* **5**, e11160, doi:10.1371/journal.pone.0011160 (2010).

113 Darcis, G. *et al.* An In-Depth Comparison of Latency-Reversing Agent Combinations in Various In Vitro and Ex Vivo HIV-1 Latency Models Identified Bryostatin-1+JQ1 and Ingenol-B+JQ1 to Potently Reactivate Viral Gene Expression. *PLoS Pathog* **11**, e1005063, doi:10.1371/journal.ppat.1005063 (2015).

114 Spina, C. A. *et al.* An in-depth comparison of latent HIV-1 reactivation in multiple cell model systems and resting CD4+ T cells from aviremic patients. *PLoS Pathog* **9**, e1003834, doi:10.1371/journal.ppat.1003834 (2013).

115 Bullen, C. K., Laird, G. M., Durand, C. M., Siliciano, J. D. & Siliciano, R. F. New ex vivo approaches distinguish effective and ineffective single agents for reversing HIV-1 latency in vivo. *Nat Med* **20**, 425-429, doi:10.1038/nm.3489 (2014).

116 Gutierrez, C. *et al.* Bryostatin-1 for latent virus reactivation in HIV-infected patients on antiretroviral therapy. *AIDS* **30**, 1385-1392, doi:10.1097/QAD.0000000000001064 (2016).

117 Webb, G. M. *et al.* The human IL-15 superagonist ALT-803 directs SIV-specific CD8(+) T cells into B-cell follicles. *Blood Adv* **2**, 76-84, doi:10.1182/bloodadvances.2017012971 (2018).

118 Jones, R. B. *et al.* A Subset of Latency-Reversing Agents Expose HIV-Infected Resting CD4+ T-Cells to Recognition by Cytotoxic T-Lymphocytes. *PLoS Pathog* **12**, e1005545, doi:10.1371/journal.ppat.1005545 (2016).

119 Watson, D. C. *et al.* Treatment with native heterodimeric IL-15 increases cytotoxic lymphocytes and reduces SHIV RNA in lymph nodes. *PLoS Pathog* **14**, e1006902, doi:10.1371/journal.ppat.1006902 (2018).

120 Vandergeeten, C. *et al.* Interleukin-7 promotes HIV persistence during antiretroviral therapy. *Blood* **121**, 4321-4329, doi:10.1182/blood-2012-11-465625 (2013).

121 Offersen, R. *et al.* A Novel Toll-Like Receptor 9 Agonist, MGN1703, Enhances HIV-1 Transcription and NK Cell-Mediated Inhibition of HIV-1-Infected Autologous CD4+ T Cells. *J Virol* **90**, 4441-4453, doi:10.1128/JVI.00222-16 (2016).

122 Vibholm, L. *et al.* Short-Course Toll-Like Receptor 9 Agonist Treatment Impacts Innate Immunity and Plasma Viremia in Individuals With Human Immunodeficiency Virus Infection. *Clin Infect Dis* **64**, 1686-1695, doi:10.1093/cid/cix201 (2017).

123 Vibholm, L. K. *et al.* Effects of 24-week Toll-like receptor 9 agonist treatment in HIV type 1+ individuals. *AIDS* **33**, 1315-1325, doi:10.1097/QAD.0000000000002213 (2019).

124 Sogaard, O. S. *et al.* The Depsipeptide Romidepsin Reverses HIV-1 Latency In Vivo. *PLoS Pathog* **11**, e1005142, doi:10.1371/journal.ppat.1005142 (2015).

125 Archin, N. M. *et al.* HIV-1 expression within resting CD4+ T cells after multiple doses of vorinostat. *J Infect Dis* **210**, 728-735, doi:10.1093/infdis/jiu155 (2014).

126 Rasmussen, T. A. *et al.* Panobinostat, a histone deacetylase inhibitor, for latent-virus reactivation in HIV-infected patients on suppressive antiretroviral therapy: a phase 1/2, single group, clinical trial. *Lancet HIV* **1**, e13-21, doi:10.1016/S2352-3018(14)70014-1 (2014).

127 Bouchat, S. *et al.* Histone methyltransferase inhibitors induce HIV-1 recovery in resting CD4(+) T cells from HIV-1-infected HAART-treated patients. *AIDS* **26**, 1473-1482, doi:10.1097/QAD.0b013e32835535f5 (2012).

128 Elliott, J. H. *et al.* Short-term administration of disulfiram for reversal of latent HIV infection: a phase 2 dose-escalation study. *Lancet HIV* **2**, e520-529, doi:10.1016/S2352-3018(15)00226-X (2015).

129 Spivak, A. M. *et al.* A pilot study assessing the safety and latency-reversing activity of disulfiram in HIV-1-infected adults on antiretroviral therapy. *Clin Infect Dis* **58**, 883-890, doi:10.1093/cid/cit813 (2014).

130 Banerjee, C. *et al.* BET bromodomain inhibition as a novel strategy for reactivation of HIV-1. *J Leukoc Biol* **92**, 1147-1154, doi:10.1189/jlb.0312165 (2012).

131 Bartholomeeusen, K., Xiang, Y., Fujinaga, K. & Peterlin, B. M. Bromodomain and extra-terminal (BET) bromodomain inhibition activate transcription via transient release of positive transcription elongation factor b (P-TEFb) from 7SK small nuclear ribonucleoprotein. *J Biol Chem* **287**, 36609-36616, doi:10.1074/jbc.M112.410746 (2012).

132 Budhiraja, S., Famiglietti, M., Bosque, A., Planelles, V. & Rice, A. P. Cyclin T1 and CDK9 T-loop phosphorylation are downregulated during establishment of HIV-1 latency in primary resting memory CD4+ T cells. *J Virol* **87**, 1211-1220, doi:10.1128/JVI.02413-12 (2013).

133 Laird, G. M. *et al.* Ex vivo analysis identifies effective HIV-1 latency-reversing drug combinations. *J Clin Invest* **125**, 1901-1912, doi:10.1172/JCI80142 (2015).

134 Zerbato, J. M., Purves, H. V., Lewin, S. R. & Rasmussen, T. A. Between a shock and a hard place: challenges and developments in HIV latency reversal. *Curr Opin Virol* **38**, 1-9, doi:10.1016/j.coviro.2019.03.004 (2019).

135 Lemieux, S. & Major, F. RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic Acids Res* **30**, 4250-4263, doi:10.1093/nar/gkf540 (2002).

136 Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* **15**, 469-479, doi:10.1038/nrg3681 (2014).

137 Strobel, E. J., Yu, A. M. & Lucks, J. B. High-throughput determination of RNA structures. *Nat Rev Genet* **19**, 615-634, doi:10.1038/s41576-018-0034-x (2018).

138 Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science* **289**, 905-920, doi:10.1126/science.289.5481.905 (2000).

139 Schluenzen, F. *et al.* Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell* **102**, 615-623, doi:10.1016/s0092-8674(00)00084-2 (2000).

140 Wimberly, B. T. *et al.* Structure of the 30S ribosomal subunit. *Nature* **407**, 327-339, doi:10.1038/35030006 (2000).

141 Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701-705, doi:10.1038/nature12894 (2014).

142 Ding, Y., Kwok, C. K., Tang, Y., Bevilacqua, P. C. & Assmann, S. M. Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. *Nat Protoc* **10**, 1050-1066, doi:10.1038/nprot.2015.064 (2015).

143 Fallmann, J. *et al.* Recent advances in RNA folding. *J Biotechnol* **261**, 97-104, doi:10.1016/j.jbiotec.2017.07.007 (2017).

144 Mathews, D. H. How to benchmark RNA secondary structure prediction accuracy. *Methods* **162-163**, 60-67, doi:10.1016/j.ymeth.2019.04.003 (2019).

145 Mathews, D. H. *et al.* Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A* **101**, 7287-7292, doi:10.1073/pnas.0401799101 (2004).

146 Peattie, D. A. & Gilbert, W. Chemical probes for higher-order structure in RNA. *Proc Natl Acad Sci U S A* **77**, 4679-4682, doi:10.1073/pnas.77.8.4679 (1980).

147 Inoue, T. & Cech, T. R. Secondary structure of the circular form of the Tetrahymena rRNA intervening sequence: a technique for RNA structure analysis using chemical probes and reverse transcriptase. *Proc Natl Acad Sci U S A* **82**, 648-652, doi:10.1073/pnas.82.3.648 (1985).

148 Tijerina, P., Mohr, S. & Russell, R. DMS footprinting of structured RNAs and RNA-protein complexes. *Nat Protoc* **2**, 2608-2623, doi:10.1038/nprot.2007.380 (2007).

149 Mohr, S. *et al.* Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **19**, 958-970, doi:10.1261/rna.039743.113 (2013).

150 Zubradt, M. *et al.* DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat Methods* **14**, 75-82, doi:10.1038/nmeth.4057 (2017).

151 Spahn, C. M. *et al.* Cryo-EM visualization of a viral internal ribosome entry site bound to human ribosomes: the IRES functions as an RNA-based translation factor. *Cell* **118**, 465-475, doi:10.1016/j.cell.2004.08.001 (2004).

152 Arragain, B. *et al.* High resolution cryo-EM structure of the helical RNA-bound Hantaan virus nucleocapsid reveals its assembly mechanisms. *Elife* **8**, doi:10.7554/eLife.43075 (2019).

153 Kirchdoerfer, R. N., Saphire, E. O. & Ward, A. B. Cryo-EM structure of the Ebola virus nucleoprotein-RNA complex. *Acta Crystallogr F Struct Biol Commun* **75**, 340-347, doi:10.1107/S2053230X19004424 (2019).

154 Gopal, A., Zhou, Z. H., Knobler, C. M. & Gelbart, W. M. Visualizing large RNA molecules in solution. *RNA* **18**, 284-299, doi:10.1261/rna.027557.111 (2012).

155 Athavale, S. S. *et al.* In vitro secondary structure of the genomic RNA of satellite tobacco mosaic virus. *PLoS One* **8**, e54384, doi:10.1371/journal.pone.0054384 (2013).

156 Garmann, R. F. *et al.* Visualizing the global secondary structure of a viral RNA genome with cryo-electron microscopy. *RNA* **21**, 877-886, doi:10.1261/rna.047506.114 (2015).

157 Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc* **127**, 4223-4231, doi:10.1021/ja043822v (2005).

158 Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods* **11**, 959-965, doi:10.1038/nmeth.3029 (2014).

159 Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A. & Weeks, K. M. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat Protoc* **10**, 1643-1669, doi:10.1038/nprot.2015.103 (2015).

160 Smola, M. J. & Weeks, K. M. In-cell RNA structure probing with SHAPE-MaP. *Nat Protoc* **13**, 1181-1195, doi:10.1038/nprot.2018.010 (2018).

161 Rausch, J. W., Sztuba-Solinska, J. & Le Grice, S. F. J. Probing the Structures of Viral RNA Regulatory Elements with SHAPE and Related Methodologies. *Front Microbiol* **8**, 2634, doi:10.3389/fmicb.2017.02634 (2017).

162 Pelletier, J. & Sonenberg, N. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature* **334**, 320-325, doi:10.1038/334320a0 (1988).

163 Mailliot, J. & Martin, F. Viral internal ribosomal entry sites: four classes for one goal. *Wiley Interdiscip Rev RNA* **9**, doi:10.1002/wrna.1458 (2018).

164 Bieleski, L. & Talbot, S. J. Kaposi's sarcoma-associated herpesvirus vCyclin open reading frame contains an internal ribosome entry site. *J Virol* **75**, 1864-1869, doi:10.1128/JVI.75.4.1864-1869.2001 (2001).

165 Yu, Y. & Alwine, J. C. 19S late mRNAs of simian virus 40 have an internal ribosome entry site upstream of the virion structural protein 3 coding sequence. *J Virol* **80**, 6553-6558, doi:10.1128/JVI.00517-06 (2006).

166 Chavez-Calvillo, G., Martin, S., Hamm, C. & Sztuba-Solinska, J. The Structure-To-Function Relationships of Gammaherpesvirus-Encoded Long Non-Coding RNAs and Their Contributions to Viral Pathogenesis. *Noncoding RNA* **4**, doi:10.3390/ncrna4040024 (2018).

167 Tolbert, M. *et al.* HnRNP A1 Alters the Structure of a Conserved Enterovirus IRES Domain to Stimulate Viral Translation. *J Mol Biol* **429**, 2841-2858, doi:10.1016/j.jmb.2017.06.007 (2017).

168 Sztuba-Solinska, J. *et al.* Kaposi's sarcoma-associated herpesvirus polyadenylated nuclear RNA: a structural scaffold for nuclear, cytoplasmic and viral proteins. *Nucleic Acids Res* **45**, 6805-6821, doi:10.1093/nar/gkx241 (2017).

169 Simon, L. M. *et al.* In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. *Nucleic Acids Res* **47**, 7003-7017, doi:10.1093/nar/gkz318 (2019).

170 Dadonaite, B. *et al.* The structure of the influenza A virus genome. *Nat Microbiol*, doi:10.1038/s41564-019-0513-7 (2019).

171 Dai, Y. *et al.* Molecular recognition of a branched peptide with HIV-1 Rev Response Element (RRE) RNA. *Bioorg Med Chem* **27**, 1759-1765, doi:10.1016/j.bmc.2019.03.016 (2019).

172 Dai, Y. *et al.* Discovery of a Branched Peptide That Recognizes the Rev Response Element (RRE) RNA and Blocks HIV-1 Replication. *J Med Chem* **61**, 9611-9620, doi:10.1021/acs.jmedchem.8b01076 (2018).

173 Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat Methods* **10**, 623-629, doi:10.1038/nmeth.2483 (2013).

174 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

175 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).

176 Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129, doi:10.1186/1471-2105-11-129 (2010).

177 Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974-1975, doi:10.1093/bioinformatics/btp250 (2009).

178 Mohammadi, P. *et al.* Dynamics of HIV latency and reactivation in a primary CD4+ T cell model. *PLoS Pathog* **10**, e1004156, doi:10.1371/journal.ppat.1004156 (2014).

179 Puglisi, J. D., Tan, R., Calnan, B. J., Frankel, A. D. & Williamson, J. R. Conformation of the TAR RNA-arginine complex by NMR spectroscopy. *Science* **257**, 76-80, doi:10.1126/science.1621097 (1992).

180 Macejak, D. G. & Sarnow, P. Internal initiation of translation mediated by the 5' leader of a cellular mRNA. *Nature* **353**, 90-94, doi:10.1038/353090a0 (1991).

181 Johannes, G., Carter, M. S., Eisen, M. B., Brown, P. O. & Sarnow, P. Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proc Natl Acad Sci U S A* **96**, 13118-13123, doi:10.1073/pnas.96.23.13118 (1999).

182 Abbink, T. E., Ooms, M., Haasnoot, P. C. & Berkhout, B. The HIV-1 leader RNA conformational switch regulates RNA dimerization but does not regulate mRNA translation. *Biochemistry* **44**, 9058-9066, doi:10.1021/bi0502588 (2005).

183 Bishop, C. M. *Recognition and Machine Learning*. (Springer, 2006).

184 Spasic, A., Assmann, S. M., Bevilacqua, P. C. & Mathews, D. H. Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Res* **46**, 314-323, doi:10.1093/nar/gkx1057 (2018).

185 Homan, P. J. *et al.* Single-molecule correlated chemical probing of RNA. *Proc Natl Acad Sci U S A* **111**, 13858-13863, doi:10.1073/pnas.1407306111 (2014).

186 Sengupta, A., Rice, G. M. & Weeks, K. M. Single-molecule correlated chemical probing reveals large-scale structural communication in the ribosome and the mechanism of the antibiotic spectinomycin in living cells. *PLoS Biol* **17**, e3000393, doi:10.1371/journal.pbio.3000393 (2019).

187 Ding, Y. & Lawrence, C. E. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res* **31**, 7280-7301, doi:10.1093/nar/gkg938 (2003).

188 Halvorsen, M., Martin, J. S., Broadaway, S. & Laederach, A. Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet* **6**, e1001074, doi:10.1371/journal.pgen.1001074 (2010).

189 Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706-709, doi:10.1038/nature12946 (2014).

190 Tian, S., Kladwang, W. & Das, R. Allosteric mechanism of the V. vulnificus adenine riboswitch resolved by four-dimensional chemical mapping. *Elife* **7**, doi:10.7554/eLife.29602 (2018).

191 Lemay, J. F., Penedo, J. C., Mulhbacher, J. & Lafontaine, D. A. Molecular basis of RNA-mediated gene regulation on the adenine riboswitch by single-molecule approaches. *Methods Mol Biol* **540**, 65-76, doi:10.1007/978-1-59745-558-9_6 (2009).

192 Lemay, J. F. *et al.* Comparative study between transcriptionally- and translationally-acting adenine riboswitches reveals key differences in riboswitch regulatory mechanisms. *PLoS Genet* **7**, e1001278, doi:10.1371/journal.pgen.1001278 (2011).

193 Zaug, A. J. & Cech, T. R. Analysis of the structure of Tetrahymena nuclear RNAs in vivo: telomerase RNA, the self-splicing rRNA intron, and U2 snRNA. *RNA* **1**, 363-374 (1995).

194 Schwartz, G. Estimating the dimension of a model. . *The Annals of Statistics* **6**, 461-464 (1978).

195 Kondo, Y., Oubridge, C., van Roon, A. M. & Nagai, K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *Elife* **4**, doi:10.7554/eLife.04986 (2015).

196 Cornilescu, G. *et al.* Structural Analysis of Multi-Helical RNAs by NMR-SAXS/WAXS: Application to the U4/U6 di-snRNA. *J Mol Biol* **428**, 777-789, doi:10.1016/j.jmb.2015.11.026 (2016).

197 Leyre, L. *et al.* Abundant HIV-infected cells in blood and tissues are rapidly cleared upon ART initiation during acute HIV infection. **12**, doi:10.1126/scitranslmed.aav3491 (2020).

198 Kwon, K. J. *et al.* Different human resting memory CD4(+) T cell subsets show similar low inducibility of latent HIV-1 proviruses. *Sci Transl Med* **12**, doi:10.1126/scitranslmed.aax6795 (2020).

199 Banga, R. *et al.* PD-1(+) and follicular helper T cells are responsible for persistent HIV-1 transcription in treated aviremic individuals. *Nat Med* **22**, 754-761, doi:10.1038/nm.4113 (2016).

200 Lorenzi, J. C. *et al.* Paired quantitative and qualitative assessment of the replication-competent HIV-1 reservoir and comparison with integrated proviral DNA. *Proc Natl Acad Sci U S A* **113**, E7908-E7916, doi:10.1073/pnas.1617789113 (2016).

201 Hosmane, N. N. *et al.* Proliferation of latently infected CD4(+) T cells carrying replication-competent HIV-1: Potential role in latent reservoir dynamics. *J Exp Med* **214**, 959-972, doi:10.1084/jem.20170193 (2017).

202 Simonetti, F. R. *et al.* Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. *Proc Natl Acad Sci U S A* **113**, 1883-1888, doi:10.1073/pnas.1522675113 (2016).

203 Musick, A. *et al.* HIV Infected T Cells Can Proliferate in vivo Without Inducing Expression of the Integrated Provirus. *Front Microbiol* **10**, 2204, doi:10.3389/fmicb.2019.02204 (2019).

204 De Scheerder, M. A. *et al.* HIV Rebound Is Predominantly Fueled by Genetically Identical Viral Expansions from Diverse Reservoirs. *Cell Host Microbe* **26**, 347-358 e347, doi:10.1016/j.chom.2019.08.003 (2019).

205 Archin, N. M. *et al.* Expression of latent HIV induced by the potent HDAC inhibitor suberoylanilide hydroxamic acid. *AIDS Res Hum Retroviruses* **25**, 207-212, doi:10.1089/aid.2008.0191 (2009).

206 Fidler, S. *et al.* Antiretroviral therapy alone versus antiretroviral therapy with a kick and kill approach, on measures of the HIV reservoir in participants with recent HIV infection (the RIVER trial): a phase 2, randomised trial. *Lancet*, doi:10.1016/S0140-6736(19)32990-3 (2020).

207 Bradley, T., Ferrari, G., Haynes, B. F., Margolis, D. M. & Browne, E. P. Single-Cell Analysis of Quiescent HIV Infection Reveals Host Transcriptional Profiles that Regulate Proviral Latency. *Cell Rep* **25**, 107-117 e103, doi:10.1016/j.celrep.2018.09.020 (2018).

208 Golumbeanu, M. *et al.* Single-Cell RNA-Seq Reveals Transcriptional Heterogeneity in Latent and Reactivated HIV-Infected Cells. *Cell Rep* **23**, 942-950, doi:10.1016/j.celrep.2018.03.102 (2018).

209 Cohn, L. B. *et al.* Clonal CD4(+) T cells in the HIV-1 latent reservoir display a distinct gene profile upon reactivation. *Nat Med* **24**, 604-609, doi:10.1038/s41591-018-0017-7 (2018).

210 Acchioni, C. *et al.* Alternate NF-kappaB-Independent Signaling Reactivation of Latent HIV-1 Provirus. *J Virol* **93**, doi:10.1128/JVI.00495-19 (2019).

211 Nixon, C. C. *et al.* Systemic HIV and SIV latency reversal via non-canonical NF-kappaB signalling in vivo. *Nature* **578**, 160-165, doi:10.1038/s41586-020-1951-3 (2020).

212 Wolschendorf, F. *et al.* Kinase control prevents HIV-1 reactivation in spite of high levels of induced NF-kappaB activity. *J Virol* **86**, 4548-4558, doi:10.1128/JVI.06726-11 (2012).

213 Duverger, A. *et al.* An AP-1 binding site in the enhancer/core element of the HIV-1 promoter controls the ability of HIV-1 to establish latent infection. *J Virol* **87**, 2264-2277, doi:10.1128/JVI.01594-12 (2013).

214 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

215 Cameron, M. J. & Kelvin, D. in *Madame Curie Bioscience Database [Internet]* (Landes Bioscience, Austin, TX, 2000-2013).

216 Turner, M. D., Nedjai, B., Hurst, T. & Pennington, D. J. Cytokines and chemokines: At the crossroads of cell signalling and inflammatory disease. *Biochim Biophys Acta* **1843**, 2563-2582, doi:10.1016/j.bbamcr.2014.05.014 (2014).

217 Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665, doi:10.1016/j.cell.2018.01.029 (2018).

218 Grau-Exposito, J. *et al.* Latency reversal agents affect differently the latent reservoir present in distinct CD4+ T subpopulations. *PLoS Pathog* **15**, e1007991, doi:10.1371/journal.ppat.1007991 (2019).

219 Roebuck, K. A., Gu, D. S. & Kagnoff, M. F. Activating protein-1 cooperates with phorbol ester activation signals to increase HIV-1 expression. *AIDS* **10**, 819-826, doi:10.1097/00002030-199607000-00004 (1996).

220 Rodriguez-Martinez, J. A., Reinke, A. W., Bhimsaria, D., Keating, A. E. & Ansari, A. Z. Combinatorial bZIP dimers display complex DNA-binding specificity landscapes. *Elife* **6**, doi:10.7554/eLife.19272 (2017).

221 Jain, J., McCaffrey, P. G., Valge-Archer, V. E. & Rao, A. Nuclear factor of activated T cells contains Fos and Jun. *Nature* **356**, 801-804, doi:10.1038/356801a0 (1992).

222 Moron-Lopez, S. *et al.* Characterization of the HIV-1 transcription profile after romidepsin administration in ART-suppressed individuals. *AIDS* **33**, 425-431, doi:10.1097/QAD.0000000000002083 (2019).

223 Ho, Y. C. *et al.* Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155**, 540-551, doi:10.1016/j.cell.2013.09.020 (2013).

224 Chen, H. C., Martinez, J. P., Zorita, E., Meyerhans, A. & Filion, G. J. Position effects influence HIV latency reversal. *Nat Struct Mol Biol* **24**, 47-54, doi:10.1038/nsmb.3328 (2017).

225 Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res* **6**, 100, doi:10.12688/f1000research.10571.2 (2017).

226 Singh, N. N., Lee, B. M. & Singh, R. N. Splicing regulation in spinal muscular atrophy by an RNA structure formed by long-distance interactions. *Ann N Y Acad Sci* **1341**, 176-187, doi:10.1111/nyas.12727 (2015).

227 Seo, J. *et al.* Oxidative Stress Triggers Body-Wide Skipping of Multiple Exons of the Spinal Muscular Atrophy Gene. *PLoS One* **11**, e0154390, doi:10.1371/journal.pone.0154390 (2016).

228 Myers, S. A. *et al.* Discovery of proteins associated with a predefined genomic locus via dCas9-APEX-mediated proximity labeling. *Nat Methods* **15**, 437-439, doi:10.1038/s41592-018-0007-1 (2018).

229 Calvanese, V., Chavez, L., Laurent, T., Ding, S. & Verdin, E. Dual-color HIV reporters trace a population of latently infected cells and enable their purification. *Virology* **446**, 283-292, doi:10.1016/j.virol.2013.07.037 (2013).