PATTERNS OF HIV INTEGRATION AND SPlicing:
WINDOWS ON MECHANISM

Karen E. Ocieja

A DISSERTATION
in
Cell and Molecular Biology

Presented to the Faculties of the University of Pennsylvania
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy
2012

Supervisor of Dissertation
Frederic D. Bushman, Ph.D.
Professor, Department of Microbiology

Graduate Group Chairperson
Daniel S. Kessler, Ph.D.
Associate Professor, Department of Cell
and Developmental Biology

Dissertation Committee
Jianxin You, Ph.D., Assistant Professor, Department of Microbiology
Paul Bates, Ph.D., Professor, Department of Microbiology
James A. Hoxie, M.D., Professor, Department of Medicine
Ronald G. Collman, M.D., Professor, Department of Medicine
PATTERNS OF HIV INTEGRATION AND SPLICING: WINDOWS ON MECHANISM

COPYRIGHT

2012

Karen E. Ocieja
DEDICATION

To my parents, who have supported, encouraged, and inspired me.
ACKNOWLEDGEMENTS

I would like to thank my advisor, Rick Bushman, for his guidance and scientific optimism, and for the resources to complete the research in this dissertation. I would also like to thank my thesis committee, Jianxin You, Jim Hoxie, Ron Collman, and Paul Bates, for their ideas, discussion, and support. I am grateful to Bob Doms, who has served as an unfailingly generous mentor, providing advice for this research as well as for my personal and professional development. I am indebted to the University of Pennsylvania Medical Scientist Training program for giving me the opportunity to do this work, and am particularly thankful for the guidance of Skip Brass and Maggie Krall.

I thank several members of the Bushman lab for teaching me to be a scientist, to think critically, and to teach clearly. In particular, I thank Keshet Ronen, Troy Brady, and Rithun Mukherjee, who instructed me in many of the methods used in this work; Keshet and Christian Hoffman for scientific discussion and argument; and Rebecca Custers-Allen, Ted Kreider, and Steven Siegel, who trained me as much as I did them. I am hugely indebted to Nirav Malani and Scott Sherrill-Mix, whose discerning and inventive computational analysis was vital to this work. Their patience in discussing, and re-discussing, data was greatly appreciated. I also thank Young Hwang and Shoshannah Roth, and Greg Peterfreund for their generous help.

I am especially grateful for the unconditional love and support of my family: Susan Pearson, Frank Ocwieja, Matthew Ocwieja, Carol Cheung, and Ethan Ocwieja, whose brilliance at the age of three inspires me to continue to learn. I also am
indebted to my friends, who have motivated and grounded me during the last several years: Drew Lynford, Joanna Sinha, Jacquelyn Mahendra, Angus Shieh, Elizabeth Siegel, Sophia Mak, Sarah Longworth, and Emerson Avery.

Finally, for their friendship, sympathy, practical jokes, and unity, I thank all the members of the Bushman lab, including Rebecca Custers-Allen, Jennifer Hwang, Christian Hoffman, Keshet Ronen, Stephanie Grunberg, Nirav Malani, Rohini Sinha, Ted Kreider, Sesh Sundararaman, Scott Sherrill-Mix, Frances Male, Emily Charlson, and Sam Minot.
ABSTRACT

PATTERNS OF HIV INTEGRATION AND SPlicing:

WINDOWS ON MECHANISM

Karen E. Ocwieja

Frederic D. Bushman

The retrovirus HIV-1 establishes lifelong infections, facilitated by integration, the covalent insertion of the viral cDNA into the host cell genome. After integration, the viral genes are expressed by host cell machinery. This dissertation focuses on patterns of HIV-1 integration site selection and of viral mRNA splicing, both of which are optimized to drive appropriate viral gene expression, and ultimately, virion production. In an effort to uncover determinants of integration site selection we investigated the role of the nuclear pore, interaction with which promotes HIV-1 integration efficiency. We found that the nuclear transport protein, Transportin-3, and the pore component, RanBP2, were important for viral targeting to chromatin rich in genes. We next asked which viral factors determine this bias. The HIV-1 CA protein has been shown to bind RanBP2. We used CA mutants to show that the preference for gene dense chromatin partially mapped to this interaction. We also sought to identify chromatin features that most strongly attract HIV-1 particles, and to this end we characterized normal distributions of HIV-1 integration sites in vi
primary human T cells at unprecedented depth. In another group of studies, we investigated viral mRNA populations and asked whether their calibration was sensitive to differences between cells. A series of RNA elements interact with host factors to regulate splicing of the HIV-1 pre-RNA transcript. We found that production or degradation of certain mRNAs differed between T cells and a human osteosarcoma cell line and among human donors. Such differences might have ties to cell cycle control. We also detected viral strain-specific splicing patterns. The patterns of integration and mRNA populations described in this dissertation reflect the importance of cellular factors in the HIV-1 replication cycle. These studies aid in the identification of therapeutic targets, while revealing much unexplored complexity in the viral steps of nuclear import, integration and mRNA processing.
TABLE OF CONTENTS

ABSTRACT ............................................................................................................................. VI

LIST OF TABLES .................................................................................................................... X

LIST OF FIGURES ................................................................................................................ XI

CHAPTER 1 – INTRODUCTION................................................................................................1

  Clinical Significance of HIV ..............................................................................................1
  Characteristics and Replication Cycle of HIV .........................................................................2
  HIV Accessory Proteins Vif, Vpr, Vpu and Nef .......................................................................5
  Nuclear Pore and Nuclear Entry of HIV .................................................................................7
  Integration and Proviral Distributions of HIV .......................................................................9
  Splicing of the HIV-1 Genome ............................................................................................13
  Deep sequencing as a probe for host-HIV interactions .......................................................20
  References ..........................................................................................................................23

CHAPTER 2 – HIV INTEGRATION TARGETING: A PATHWAY INVOLVING
TRANSPORTIN-3 AND RAN-BINDING PROTEIN 2 .............................................................. 32

  Abstract ............................................................................................................................32
  Introduction ......................................................................................................................33
  Materials and Methods ......................................................................................................36
  Results ..................................................................................................................................45
  Discussion ..........................................................................................................................63
  Acknowledgements ..........................................................................................................67
  References ..........................................................................................................................69

CHAPTER 3 – HIV CAPSID INTERACTIONS WITH CYCLOPHILIN-A AND RAN-BINDING
PROTEIN 2 INFLUENCE NUCLEAR IMPORT AND INTEGRATION PATHWAYS .............. 74

  Abstract ............................................................................................................................74
  Introduction ......................................................................................................................75
  Methods ............................................................................................................................78
  Results ..................................................................................................................................81
  Discussion ..........................................................................................................................91
  Acknowledgements ..........................................................................................................94
  References ..........................................................................................................................95

CHAPTER 4 – INTEGRATION SITE SELECTION BY A HIV89.6, A LOW PASSAGE ISOLATE,
IN PRIMARY HUMAN T CELLS ............................................................................................ 97

  Abstract ............................................................................................................................97
  Introduction ......................................................................................................................97
  Methods ............................................................................................................................100
  Results ................................................................................................................................107
  Discussion ........................................................................................................................120
  Acknowledgements ..........................................................................................................122
  References ........................................................................................................................123
CHAPTER 5 – DYNAMIC REGULATION OF HIV-1 MRNA POPULATIONS ANALYZED BY
SINGLE MOLECULE ENRICHMENT AND LONG READ SEQUENCING .............................126

Abstract ................................................................................................................................. 126
Introduction ............................................................................................................................ 127
Methods ................................................................................................................................. 129
Results ................................................................................................................................. 142
Discussion ............................................................................................................................ 160
Acknowledgements ............................................................................................................. 164
References .......................................................................................................................... 164

CHAPTER 6 – DYNAMIC HIV-1 MESSAGE POPULATIONS AND CELL CYCLE ARREST ...169

Abstract ................................................................................................................................. 169
Introduction ............................................................................................................................ 169
Methods ................................................................................................................................. 172
Results ................................................................................................................................. 175
Discussion ............................................................................................................................ 181
Acknowledgements ............................................................................................................. 184
References .......................................................................................................................... 184

CHAPTER 7 – CONCLUSIONS AND FUTURE DIRECTIONS ..............................................188

References .......................................................................................................................... 198

APPENDIX ............................................................................................................................200

 Appendix 1 - Guide to genomic heat map summaries of proviral distributions .............. 200
 Appendix 2 – Curriculum Vitae ........................................................................................... 217
### LIST OF TABLES

| Table 2-1 | siRNA oligonucleotides used in this study | 38 |
| Table 2-2 | DNA oligonucleotides used in this study | 42 |
| Table 2-3 | Integration site datasets used in this study | 44 |

| Table 3-1 | DNA oligonucleotides used in this study | 79 |
| Table 3-2 | Integration site datasets used in this study | 80 |
| Table 3-3 | Summary of HIV-1 CA mutant interactions with host factors | 83 |

| Table 4-1 | DNA oligonucleotides used in this study | 104 |
| Table 4-2 | Integration site datasets used in this study | 106 |

| Table 5-1 | DNA oligonucleotides used in this study | 133 |
| Table 5-2 | Human T cell donor information | 143 |
| Table 5-3 | HIV-189.6 splice donors and acceptors | 146 |
| Table 5-4 | HIV-189.6 splice junction counts in sequences from T cells | 148 |
| Table 5-5 | Ratios of HIV-189.6 transcripts in T cells and HOS cells | 151 |
| Table 5-6 | ORFs and amino acid sequences of Tat^8c and Ref | 155 |

| Table 6-1 | DNA oligonucleotides used in this study | 174 |
| Table 6-2 | Cell cycle inhibitors used in this study | 177 |
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure no.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1-1</td>
<td>HIV-1 components and replication</td>
<td>6</td>
</tr>
<tr>
<td>Figure 1-2</td>
<td>Nuclear pore</td>
<td>8</td>
</tr>
<tr>
<td>Figure 1-3</td>
<td>The HIV-1 integration reaction</td>
<td>10</td>
</tr>
<tr>
<td>Figure 1-4</td>
<td>HIV-1 genome, transcripts, and splicing regulation</td>
<td>14</td>
</tr>
<tr>
<td>Figure 1-5</td>
<td>Eukaryotic Splicing</td>
<td>15</td>
</tr>
<tr>
<td>Figure 2-1</td>
<td>Cell viability after siRNA transfection</td>
<td>47</td>
</tr>
<tr>
<td>Figure 2-2</td>
<td>siRNA depletion of targeted factors</td>
<td>48</td>
</tr>
<tr>
<td>Figure 2-3</td>
<td>Transportin-3 is required for infection of macrophages</td>
<td>49</td>
</tr>
<tr>
<td>Figure 2-4</td>
<td>Effects of siRNA treatments on HIV integration in gene dense regions</td>
<td>52</td>
</tr>
<tr>
<td>Figure 2-5</td>
<td>Transfection of a Transportin-3 allele insensitive to TNPO3 si4 restores protein expression, HIV infectivity, and partially restores wild-type HIV integration site distributions</td>
<td>54</td>
</tr>
<tr>
<td>Figure 2-6</td>
<td>Effects of Transportin-3 depletion on integration near multiple chromosomal features</td>
<td>56</td>
</tr>
<tr>
<td>Figure 2-7</td>
<td>Depletion of Transportin-3 does not alter MLV integration targeting</td>
<td>57</td>
</tr>
<tr>
<td>Figure 2-8</td>
<td>MLV infection and integration site distributions after siRNA treatment targeting SNW1 and ANAPC2</td>
<td>58</td>
</tr>
<tr>
<td>Figure 2-9</td>
<td>A chimeric derivative of HIV containing MLV gag (HIVmGag) shows reduced integration frequency in gene dense regions</td>
<td>61</td>
</tr>
<tr>
<td>Figure 2-10</td>
<td>HIV and MLV integration patterns at transcription start sites are unaffected by knockdown of TNPO3 or RANBP2</td>
<td>63</td>
</tr>
<tr>
<td>Figure 2-11</td>
<td>Model for coupling nuclear import and integration targeting</td>
<td>68</td>
</tr>
<tr>
<td>Figure 3-1</td>
<td>Crystal structure of HIV-1 CA hexamer</td>
<td>83</td>
</tr>
<tr>
<td>Figure 3-2</td>
<td>Effects of CA mutants and knockdowns on integration frequency near multiple chromosomal features</td>
<td>87</td>
</tr>
<tr>
<td>Figure 3-3</td>
<td>Effects of knockdowns on integration site distributions of CA mutants</td>
<td>88</td>
</tr>
<tr>
<td>Figure 3-4</td>
<td>Effects of capsid mutations are dominant to that of Cyclosporine treatment</td>
<td>90</td>
</tr>
<tr>
<td>Figure 3-5</td>
<td>Model for multiple pathways into nucleus</td>
<td>93</td>
</tr>
<tr>
<td>Figure 4-1</td>
<td>Minimizing PCR cross-contamination in amplifying integration sites</td>
<td>103</td>
</tr>
<tr>
<td>Figure 4-2</td>
<td>Integration sites, hot sites, and hot regions</td>
<td>108</td>
</tr>
<tr>
<td>Figure 4-3</td>
<td>HIV-189.6 Integration sites compared to distributions of chromosomal features</td>
<td>111</td>
</tr>
<tr>
<td>Figure 4-4</td>
<td>Integration intensity relative to density of epigenetic markers</td>
<td>112</td>
</tr>
<tr>
<td>Figure 4-5</td>
<td>Distribution of hot sites relative to density of chromosomal and epigenetic markers</td>
<td>116</td>
</tr>
<tr>
<td>Figure 4-6</td>
<td>Favored local sequence at HIV-1 integration sites</td>
<td>118</td>
</tr>
<tr>
<td>Figure 5-1</td>
<td>Locations of DNA oligonucleotides used in this study with respect to genes and exons of the HIV-1(^{189.6}) genome</td>
<td>132</td>
</tr>
<tr>
<td>Figure 5-2</td>
<td>Mapping the splice donors and acceptors of HIV-1(^{189.6})</td>
<td>145</td>
</tr>
<tr>
<td>Figure 5-3</td>
<td>Spliced transcripts produced from HIV-1(^{189.6})</td>
<td>150</td>
</tr>
<tr>
<td>Figure 5-4</td>
<td>Novel transcripts utilizing acceptor A8c</td>
<td>154</td>
</tr>
<tr>
<td>Figure 5-5</td>
<td>Temporal, cell-type, and donor variability in accumulation of HIV-1 messages</td>
<td>159</td>
</tr>
<tr>
<td>Figure 6-1</td>
<td>Increased relative abundance of HIV-189.6 mRNAs encoding Tat with time post infection in HOS-CD4-CCR5 cells</td>
<td>176</td>
</tr>
<tr>
<td>Figure 6-2</td>
<td>Treatment of infected cells with Nocodazole results in increased relative levels of tat mRNAs</td>
<td>168</td>
</tr>
<tr>
<td>Figure 6-3</td>
<td>Role of Vpr in calibrating HIV-1 mRNA populations in HOS and CD4+ T cells</td>
<td>180</td>
</tr>
</tbody>
</table>
CHAPTER 1 – INTRODUCTION

Clinical Significance of HIV

In the late 1970s and early 1980s, a spike in deadly opportunistic infections and unusual cancers in young individuals, especially among populations of men who have sex with men and intravenous drug users, heralded the beginning of the acquired immunodeficiency syndrome (AIDS) epidemic. In 1983, the etiological agent for this disease was identified as the human immunodeficiency virus (HIV) [1,2]. HIV is spread by sexual, percutaneous, and perinatal routes and attacks human macrophages and T cells expressing the receptor CD4. After an asymptomatic period averaging several years during which CD4+ T cells decline, untreated infected individuals are left with too few of these cells to defend against pathogens normally controlled by the immune system. More than 30 years since it began, the AIDS pandemic is thought to have claimed over 46 million lives, with an estimated 1.8 million AIDS related deaths and 34 million people living with HIV in 2010 [3].

Fortunately, research has led to the development of over 30 pharmaceutical inhibitors of viral replication, which, when taken in cocktails of three or more drugs, control the virus [4], profoundly improve morbidity and mortality, and limit the chances for transmission [5]. However, these drugs must be taken indefinitely, and due to the rapid mutation rate of the virus, drug resistant strains often emerge, necessitating medication changes. Moreover, high costs prohibit widespread
treatment in the developing world. While early evidence suggests gene therapy approaches could take the place of medication, no cure or vaccine is yet available [6,7]. Therefore we continue to dissect the replication cycle of HIV with the hope of identifying steps sensitive to novel therapeutic intervention.

Meanwhile, the gene therapy field has begun to exploit HIV and related viruses as vectors that can efficiently deliver genetic material to cells in order to restore expression of a required protein or stably produce a therapeutic agent within a patient [8,9]. Understanding viral processes is key to minimizing risk and maximizing efficacy in the development of such treatments. In both the gene therapy and microbiology fields there has been particular interest in HIV’s interactions with and dependence on host-derived factors during its replication cycle. The research described in this dissertation provides insight into the viral sensitivity to host nuclear machinery, which might be targeted by anti-retroviral agents or modulated for better control of gene therapy vectors.

**Characteristics and Replication Cycle of HIV**

HIV is a member of the *Retroviridae* family of spherical enveloped RNA viruses defined by two important steps of their replication cycles: reverse transcription of the RNA genome into a cDNA copy, and integration of that cDNA into the host genome. The *Retroviridae* are broken into two sub-families, *spumaretroviridae* and *orthoretroviridae*. The latter is further subdivided into six genera grouped by sequence similarity [10]. HIV-1 (responsible for most HIV
disease) and the related HIV-2 and simian immunodeficiency viruses (SIV) are lentiviruses, which typically cause chronic persistent infections in mammalian hosts [11].

The mature HIV-1 virion, depicted in Figure 1-1A, is enveloped by a host-cell-derived lipid bilayer studded with viral viral Env glycoproteins gp120 (SU) and gp41 (TM) and lined with viral matrix protein (MA). Within this envelope lies the viral core, which contains two copies of the single stranded HIV-1 RNA genome, stabilized by viral nucleocapsid (NC), and several host- and virus-derived molecules including viral enzymes integrase (IN) and reverse transcriptase (RT). These components are packaged within a protein cage, or capsid, made up of viral capsid protein (CA) [12]. A lattice of CA hexamers is interrupted with pentamers to give the capsid a characteristic conical shape [13,14].

The HIV replication cycle begins at viral entry, which is thought to occur either at the cell surface or within endosomes [11] (Fig. 1-1B, [15]). Interactions with host proteins found on the surface of the target cell – CD4 and a chemokine receptor CCR5 or sometimes CXCR4 – induce conformational changes in the Env glycoproteins that lead to fusion of the viral lipid envelope with the host cell membrane and viral entry [11,16-18]. This step releases the capsid-bound viral core into the cytoplasm. Capsid disassembly, known as uncoating, is intrinsically linked to the steps of reverse transcription and formation of the pre-integration complex (PIC), but the timing, location and mechanism of this process are poorly defined. Both stabilizing and destabilizing mutations in CA abrogate reverse
transcription and efficient infection [12,19,20]. Moreover, Trim5α from some primates combats HIV-1 infection by binding and destabilizing capsid [21-23], while human Cyclophilin-A is recruited to and modifies CA, possibly protecting the capsid from innate immune factors [24,25]. Viral reverse transcriptase (RT) catalyzes the creation of a cDNA copy of the viral genome probably in the context of a partially disassembled capsid [12]. This step requires host derived nucleotides and a tRNA primer and may occur as the core factors travel along the cellular cytoskeleton towards the nucleus [11]. Next, further conformational changes to the core yield the pre-integration complex (PIC), a nucleoprotein structure that is transported into the nucleus where IN catalyzes the integration of the viral cDNA into the host chromatin. HIV PICs isolated from the cytoplasm have been shown to contain viral IN, MA, RT, NC, and Vpr [26]. It is unclear whether viral CA is also included in these complexes [27].

The integrated viral genome, known as the provirus, functions as an endogenous gene and is transcribed and translated by the host machinery. The HIV Tat protein serves as an elongation factor to enhance transcription of the viral genome [28-30]. Host factors splice viral transcripts into several mRNAs encoding the 9 viral proteins and polyproteins, and viral Rev aids in exporting these from the nucleus [31,32]. Env is translated in the host cell endoplasmic reticulum, modified and processed in the Golgi and then transported to the cellular membrane where the transmembrane subunit interacts with lipid microdomains or “rafts” high in cholesterol. The polyproteins Gag and Gag-Pol also associate with lipid rafts and
Env [11]. Packaging of the unspliced HIV RNA genome requires sequences in the 5’ UTR. Watson Crick base-pairing at the dimerization initiation site (DIS) enables the genome to form dimers which then interact with the nucleocapsid (NC) domain of Gag via the RNA loop $\psi$ [11,33].

Assembly at the plasma membrane leads to budding of the virion from the cell. Coincident conformational changes activate the viral protease PR, part of the Pol polyprotein, to cleave Gag into its domains MA, CA and NC and Pol into PR, RT and IN. This enables virion maturation, including assembly of the capsid and condensation of the viral RNA [11]. Additional viral and host proteins have been identified in mature virions including the tRNAs that prime the RT reaction [11,34].

**HIV Accessory Proteins Vif, Vpr, Vpu and Nef**

In addition to the essential viral components discussed above, HIV produces accessory proteins, Vif, Vpr, Vpu, and Nef, which largely oppose host defenses. Nef relocates MHC I molecules and T-cell receptors (TCR-CD3) from the cell surface to prevent interactions of the infected T cell with the immune system [35], and, with the help of Vpu, reduces surface CD4 to facilitate budding, prevent superinfection of the cell, and compromise the ability of T cells to respond to immunogenic peptides [36]. HIV-1 Vif and Vpu combat the innate immune factors APOBEC3G and tetherin respectively. Vpr is less well understood. It is required for replication in macrophages [37] and induces cell cycle arrest at the G2 phase in T cells, potentially to benefit the virus or impair T cell function [35].
Figure 1-1 HIV-1 components and replication. (A) The HIV-1 particle. (B) The HIV-1 replication cycle. Adapted from Ronen [15].
Nuclear Pore and Nuclear Entry of HIV

The nuclear pore is a 60 MDa structure spanning the eukaryotic nuclear envelope that allows selective passage of material between the cytoplasm and nucleus [38] (Fig. 1-2). It is made up of structural nucleoporin (NUP) proteins, many of which contain phenylalanine-glycine (FG) repeat domains that make required contacts with traversing molecules [39]. Material is shuttled through the pore by soluble receptors called karyopherins with directionality dictated by Ran, a GTPase that associates with import complexes in the GDP-bound form [40]. Ran-binding protein 2 (RanBP2, also known as Nup358) interacts with the Ran-GTPase, its activating protein, and several other components of nuclear transport machinery at the cytoplasmic face of the nuclear pore [41,42].

HIV and other lentiviruses are unique among retroviruses in that they can productively infect non-cycling cells such as macrophages. Thus, they must be able to access the nucleus of these cells, and do so almost certainly by means of the nuclear pore [43-45]. Other orthoretroviruses, like the gammaretrovirus murine leukemia virus (MLV), only accumulate within the nucleus after nuclear envelope breakdown during mitosis [44,46]. Support for the belief that HIV crosses the nuclear pore came from three siRNA screens conducted in 2008. These screens were designed to identify host cofactors for HIV infection, and reported requirements for several pore factors including RanBP2 [47-50]. In addition, Krogan and colleagues have detected binding interactions between viral proteins and multiple elements of the nuclear transport machinery [51].
**Figure 1-2 - Nuclear Pore.** Karyopherins like Transportin-3 interact with the phenylalanine-glycine (FG) repeats of the nuclear pore structural proteins (nucleoporins, NUPs) and shuttle cargo through the nuclear pore. Ran-GTP binding in the nucleus displaces imported cargo and allows export of karyopherins. The Ran-guanine nucleotide exchange factor (GEF), RCC1, associates with chromatin. The Ran-GTPase activating protein (GAP) and the Ran-binding protein RANBP2 are located at the cytoplasmic filaments.
Several viral components of the HIV PIC have been proposed to mediate import through the nuclear pore, including MA [52], Vpr [53,54], IN [55-57], and a cis-acting 3 stranded viral cDNA structure [58]. PIC-bound tRNAs have also been proposed to play this role [59]. Likewise, several host karyopherins have been proposed to bind these elements and import the PIC [54-57]. At the time this dissertation research was begun, Transportin-3 (TNPO3 or TRN-SR2), which was identified in two of the siRNA cofactor screens [47,49], had recently been demonstrated as required for HIV infection in macrophages, import of GFP-IN into the nucleus of intact HeLaP4 cells [57], and nuclear accumulation of viral cDNA in 293T cells [49]. In the cell, TNPO3 imports a family of serine/arginine-rich pre-mRNA splicing factors (SR proteins) [60]. It localizes both to the nuclear pore and to subnuclear speckles marked by the presence of splicing factor SC-35 [61].

Integration and Proviral Distributions of HIV

Chapters 2 and 3 of this dissertation focus on HIV integration into cellular chromatin. After entering the nucleus, the PIC integrates in three steps depicted in Figure 1-3: viral IN processes the termini of the linear viral cDNA leaving recessed 3’ hydroxyl groups; it then joins the viral cDNA ends to staggered sites in the target DNA; and finally, the resulting single stranded regions flanking the integration site are repaired, probably by host factors [62-65]. The covalent joining of the HIV-1 cDNA to the host genome enables the virus to create permanent latent reservoirs within host cells that cannot be eliminated by the host immune system or with
pharmaceutical antiretroviral approaches [66]. Thus, integration is major barrier to eradication of HIV. One inhibitor of HIV integration, Raltegravir, is FDA approved, while additional small molecules are in trials; however, these only prevent new integration events in the context of a lifelong-infection. Understanding retroviral integration is of additional importance within the gene-therapy field, where stable expression of the delivered genetic material is required but insertional disruption of host gene expression must be avoided [67].

Figure 1-3 - The HIV-1 Integration reaction. (A) The pre-integration complex contains reverse transcribed viral cDNA (red) with ends bound by viral integrase (IN). Filled circles represent 5’ ends of the DNA. (B) Terminal processing. IN removes 2 nucleotides from the 3’ ends of the viral cDNA exposing 5’ overhangs and leaving 3’ hydroxyl groups. (C) Strand transfer. HIV-1 IN catalyzes the nucleophilic attack by the viral 3’-OH groups on the target host DNA phosphate backbone at two positions separated by 5 base pairs. (D) The HIV-1 integration intermediate contains 5 unpaired bases at either end of the viral DNA. (E) Gap repair. The unpaired gaps are filled in, likely by host DNA damage repair enzymes, which also cleave the viral 5’ overhangs. (F) The integrated provirus is flanked on either end by direct repeats of 5 host base pairs. Adapted from Cluffi and Bushman [75].
While IN alone confers only slight sequence specificity to integration sites [68-70], HIV integration is biased, integrating into bent DNA on nucleosomes [71,72], with strong preferences for active transcription units in vivo [73]. Several genomic features are associated with active transcription units including high G/C content, high gene density, high CpG island density, short introns, and certain epigenetic modifications [71,74-77]. These features also appear to be targets for HIV integration in all cell types tested [71,74,78-80], although slight variability has been observed. In macrophages, for example, viral integration targeting biases are similar but weakened [78].

Two models might explain the preference for integration in active transcription units. First, active transcription units are more likely to be found in more loosely packed euchromatin, and therefore may be more accessible to the HIV PIC. Alternatively (or additionally), the tethering model holds that the PIC interacts with chromatin binding proteins that may target it to specific regions of the chromosomes or to specific epigenetic markers. The accessibility model is opposed by the observation that retroviruses of other genera show different proviral distributions in the same cell types. For example, the gammaretrovirus MLV favors transcription start sites while proviral distributions of ASLV, an alpharetrovirus, are more random [81]. Therefore, the tethering model is favored.

For HIV, PSIP1/LEDGF/p75 (LEDGF) is thought to be the tether largely responsible for reported proviral distributions. In cells depleted of LEDGF, HIV provirus distributions become much less biased [75,82]. LEDGF has been shown to
bind HIV IN [83] and tether the PIC to preferred host chromatin [84,85]. It acts as a cofactor for HIV-1 IN in vitro [83,86,87] and is required for efficient HIV integration in vivo [88,89]. However, even in the presence of LEDGF, in vitro reactions are inefficient: in a substantial fraction of end joining reactions only one of the two DNA ends is joined to the target [86,87]. Moreover, in mouse embryonic fibroblasts null for LEDGF/p75, HIV retains a slight preference for integration in transcription units, and shows new preferences for CpG islands and gene promoters [82]. This data suggests there might be other host factors that assist the integration reaction.

Prior to the beginning of research for this dissertation, siRNA screens had identified components of the nuclear pore that surprisingly acted as co-factors promoting HIV integration efficiency in addition to those that were required for nuclear localization. Preliminary work by members of the Bushman laboratory showed that HIV integration site selection might also be sensitive to depletion of several nuclear factors including components of the nuclear import machinery such as Transportin-3. These observations suggested that the steps of nuclear entry and viral integration might be coupled, and that passage through the nuclear pore may afford the PIC access to additional cofactors for HIV-1 IN or to particular regions of chromatin. It remained to be determined which nuclear host factors were key in integration targeting and how they interacted with the virus.

In Chapter 2, we investigate the role of nuclear pore factors, Transportin-3 and RanBP2, in modulating integration targeting. We find that these proteins as well as HIV Gag are required for proper targeting to genomic regions rich in genes
and associated features. In Chapter 3, we present data suggesting more specifically that direct interactions between RanBP2 and CA are required to achieve normal proviral distributions. Finally in Chapter 4, we use available data on genomic features targeted for HIV integration to model integration sites in silico. We find that our model fails to fully describe the observed proviral distributions of HIV in CD4+ T cells, suggesting novel targeting mechanisms are at play and yet to be discovered.

**Splicing of the HIV-1 Genome**

The remaining experiments described in this dissertation investigate regulation of the populations of HIV mRNAs produced during infection. The ~10kb HIV proviral genome is flanked by two directly repeated sequences, the long terminal repeats (LTRs), which have three components, U3, R, and U5 (Fig. 1-4A). Only one initial pre-mRNA can be transcribed from the genome, starting at the transcription start site located at the junction of U3 and R at the 5′ end of the genome and terminating at the polyadenylation site located at the junction of the 3′ R and U5. However, alternative splicing of that RNA produces an array of mRNA messages encoding the 9 proteins and polyproteins of HIV [90,91] (Fig. 1-4B). The steady state ratios of these messages, established by splicing and differential degradation of the mRNAs, helps to determine the relative expression of the different viral proteins.
Figure 1-4 - HIV-1 Genome, Transcripts, and Splicing Regulation. (A) HIV-1 Genome. Major splice donors (D) and splice acceptors (A) are labeled in black with cryptic or poorly conserved sites in gray. Numbering is according to Purcell and Martin [91]. RNA elements are shown in white. (LTR, long terminal repeat; TSS, transcription start site; DIS, dimerization initiation sequence; ψ, packaging signal; RRE, Rev response element; PAS, polyadenylation site). (B) HIV-1 transcripts. Exons 2 and 3 are variably included where shown. Proteins encoded by each transcript are indicated at left, transcript classes at right. (C) Cis acting splicing regulatory elements of HIV-1. Splicing enhancers are shown in green, suppressors in red with respect to HIV-1 exons. Host-derived trans-acting factors that interact with HIV-1 elements are shown where known. Host factors compete for binding at ESE2 and ESS2. Adapted from [90].
Figure 1-5 - Eukaryotic Splicing. (A) The spliceosome and the splicing reaction. i. Components of the spliceosome (purple) recognize the splice donor (SD), branch-point sequence (BPS) and splice acceptor (SA) by base pairing and other interactions. SF1 is replaced by U2 at the BPS. The spliceosome catalyzes two transesterification reactions to remove the intron. ii. First the 2′ hydroxyl of the BPS adenine nucleotide attacks the phosphodiester bond at the SD. iii. Second, the new free hydroxyl at the SD attacks the phosphodiester bond at SA. iv. These transesterifications join the two exons and release a looped (lariat) intron. Adapted from Alberts et al. [92]. (B) Consensus sequences for splicing [96,97]. The branch point is usually much closer to the SA than the SD.
The eukaryotic host splicing machinery removes the HIV introns. The small nuclear ribonucleoprotein (snRNP), U1, binds to the 3’ end of the 5’ exon, the splice donor (SD), U2AF recognizes the splice acceptor (SA) at the 5’ end of the 3’ exon, and branch point binding protein (SF1/BBP) and the U2 snRNP sequentially bind the intronic branch point sequence (BPS). These factors as well as several additional snRNPs interact and rearrange within the dynamic “spliceosome” to catalyze the splicing transesterification reactions [92] (Fig. 1-5A).

The frequencies of use – or strengths – of the SD, BPS, and SA are determined by how well they are recognized by components of the spliceosome (usually by base-pairing), modulated by additional interactions between regulatory cis-acting sequences and trans-acting factors [93,94]. However, the sequences at eukaryotic SDs, branch points, and SAs are quite variable (Fig. 1-5B). The most tightly conserved features are the 5’-GU-3’ at the +1 and +2 locations following the SD, the BPS adenosine, and a poly-pyrimidine tract of variable length followed by a 5’-AG-3’ at the -2 and -1 positions upstream of the SA, but splicing has been reported at sites departing from these conservations [95-98]. Splicing at the 3’ end of an exon often also promotes splicing at its 3’ end, a phenomenon known as exon definition [92].

Most HIV-1 strains use four conserved SDs and eight conserved SAs to produce at least 47 different mRNAs shown in Figure 1-4B. The unspliced mRNA serves as the genome as well as transcripts for both the Gag and Gag/Pol polyproteins due to translational frameshifting [99]. Partially spliced transcripts, ~4kb in length, retain the second major intron to encode Vif, Vpr, a truncated 72
amino acid form of Tat that retains transactivation activity [100], and Env and Vpu in bicistronic messages. The completely spliced transcripts, ~2kb in length, encode Tat, Rev, and Nef. Throughout this dissertation, we will label SDs and SAs of HIV-1 “D” and “A” respectively, followed their numbers as indicated in Figure 1-4.

There is considerable unexplained redundancy in the system with several transcripts encoding most proteins. Exons 2 and 3 are noncoding, and may be included or omitted in most spliced transcripts, with poorly understood consequences [90]. Further complexity is added by rare “cryptic” splice acceptors and donors, which yield additional mRNAs in some HIV-1 strains. For example, the inclusion of a small cryptic exon within env (between A6 and D5), creates a message encoding a Tat-Rev-Env fusion protein (Tev) in HIV-1HXB2 that is absent in HIV-1NL4-3 [91,101,102]. Lutzelberger and colleagues reported that the SD sequence of a small infrequently incorporated exon within pol (A1a-D1a) stabilizes unspliced RNAs [103]. Splicing to two out-of-frame acceptors just upstream of A7 has been demonstrated in the context of a mutant A7 [91,104]. Finally, splicing at a set of acceptors within the 3’ LTR had been observed to yield novel messages in patient samples prior to work described in this dissertation; however, none of these could be translated to a protein of significant length [105,106]. It has been suggested that evolution of cryptic splice sites might be evolutionarily beneficial to HIV, enabling it to generate novel proteins and/or enabling escape from compromising mutations.

A complex array of cis-acting exonic and intronic splicing enhancers (ESEs and ISEs) and silencers (ESSs and ISSs) interact with trans-acting host factors to
regulate the relative abundance of the HIV-1 mRNAs produced [90] (Fig. 1-4C). Host heteronuclear ribonucleoproteins (hnRNPs) generally inhibit splicing while a family of serine and arginine-rich proteins (SR proteins) typically attract spliceosome components to nearby HIV splice sites [48]. HIV-1 SAs are weak compared to typical cellular acceptors, due to use of non-consensus branch point sequences [107], short pyrimidine tracts, and poor matches to the SA consensus [108], likely facilitating regulation.

Careful studies have characterized the balanced proportions of mRNAs produced from HIV-1 and have shown that mRNAs spliced at the acceptor A5, encoding Env/Vpu and Nef, generally predominate [91,109]. Interactions between a guanosine-adenosine rich ESE in exon 5 (GAR ESE) and the SR protein SF2/ASF – and, to a lesser degree, SRp40 – increase splicing at A5, despite the intrinsic weakness of the acceptor sequence [110,111]. Mutation of the strong donor D4 decreases upstream splicing to A5 as well as A4a, b and c, an outcome rescued by mutation of U1 to match the mutant D4, suggesting exon definition may also promote splicing at these sites [110]. By contrast, splicing at A3, a stronger acceptor that produces tat transcripts, is inhibited by binding of hnRNPs to several ESS elements in exon 4 [90], such that these transcripts contribute only a minor proportion of HIV messages [91].

The importance of maintaining appropriate ratios among HIV-1 messages is highlighted by the observation that, while evolutionarily distinct Group M (main) and O (outlier) HIV-1 viruses have different sequences regulating splicing, both
ultimately achieve similar ratios of transcripts [112,113]. Indeed perturbing this system is often detrimental to the virus. For example, mutating two inhibitory elements that reduce splicing at A3 increased production of tat mRNAs and decreased viral production 10-fold [90]. Similarly, disruption of an ESS (ESSV) that opposes splicing at A2 caused a 10- to 20-fold reduction in virion production, likely due to decreased unspliced viral RNA and Gag [113]. Altering the intrinsic strengths of HIV-1 splice sites or perturbing host-derived regulators also has been shown to impair HIV-1 replication [47-50,114-120] [121].

Our understanding of HIV-1 splicing regulation and message populations derives mostly from studies of lab-adapted viral strains in cell lines; however, splicing patterns are known to differ among HIV-1 strains [91,101,102], among endogenous human genes in a tissue-specific manner [94], and between human individuals [122,123]. Three findings underscored the importance of studying HIV-1 splicing in the favored host target cells. First, Lynch and colleagues identified >150 genes whose transcripts showed different splicing patterns between resting and active T cells [124]. HIV predominantly infects and replicates in activated T cells, while latency is established in resting cells [125], but the role of activation has not been investigated in the context of HIV splicing. Second, Sonza and colleagues found that, in infected macrophages, decreases in HIV production correlated with a decrease in tat mRNAs and changes in host trans-acting splicing regulators, suggesting that HIV might modulate host splicing factors in specific cells [115]. Finally, Silver and coworkers showed that depletion of cell cycle factors induced
splicing of the pro-apoptotic forms of Bcl-x and Mcl1 [126]. This raises the question of whether cell cycle arrest induced by HIV Vpr similarly alters the splicing milieu, affecting populations of viral and/or host mRNAs. While Vpr has been suggested to decrease HIV splicing in vitro, its role in vivo is unclear [127,128].

Prior to this dissertation, little work had been done to characterize HIV-1 mRNA populations produced by clinical viral strains in relevant primary cells. It remained unclear

1. whether HIV-1 message populations described in cell lines accurately represented those produced in activated T cells
2. whether human polymorphisms affected HIV-1 splicing patterns
3. with what frequency and biological significance did cryptic splicing events occur in T cells
4. whether HIV modulates the splicing machinery of the infected cell.

These questions motivated the work described in the second half of this dissertation. In Chapter 5 we use high-throughput sequencing to exhaustively tabulate the transcripts produced in T cells from various human donors and in a human-osteosarcoma cell line. In Chapter 6, we consider the effect of HIV infection on host-mediated splicing.

**Deep sequencing as a probe for host-HIV interactions**

This dissertation centers on the use of high-throughput sequencing techniques to identify proviral and viral mRNA patterns that reflect interactions
with the host. Because both integration and splicing exhibit immense variability — the entire human genome serves as a substrate for HIV integration, and hundreds of viral exon combinations are possible — the large datasets obtained using deep sequencing techniques are vital to make statistically meaningful observations. Moreover, in the case of splicing, multiple mRNAs have the same length, therefore gel-based assays that identify RNAs by size are not adequate, and any novelty must be identified by sequencing.

Various high-throughput sequencing technologies are available, and the choice of which to use depends on several factors including price, speed, starting material, and the sequence read length and quality (lack of errors) required for the study [129]. In the case of HIV integration, we identify proviral locations based on the human genomic sequence immediately adjacent to an LTR. Roche 454 Sequencing [130] [www.454.com] provides mid-sized reads with relatively few errors (up to approximately 400 bp with 2-3% error omitting homopolymer length errors, internal data), enabling unequivocal assignment of a given sequence to a single location in the human genome. In this technology, single molecules of DNA are amplified on individual beads. Sequencing is then carried out by synthesis, with each incorporation read out as luciferase signal.

For studies of splicing, if one is interested in describing the full structure of mRNAs, it is imperative that a single sequence read span several exons. Pacific Biosciences developed a next-generation sequencing technology that provides the long-read sequences required for such applications [131,132]. Templates are
transformed into single stranded circles (SMRTbell) by the addition of terminal
hairpin linkers and loaded with linker primers onto individual DNA polymerase
enzymes, each of which is fixed at the bottom of a small volume (20zL) well. As each
labeled base is incorporated, it is immobilized at the polymerase and detected by
live fluorescence microscopy. Because signals are detected in real time, the natural
processivity of polymerase is utilized and long sequences of up to ~10kb (personal
communication) can be read quickly. While error rate is higher than in 454
sequencing, the SMRTbell topology mitigates this by enabling rolling circle
polymerization - multiple rounds of sequence from a single template are condensed
into a “circular consensus sequence” (CCS) with fewer errors.

In this dissertation, we use Roche 454 pyrosequencing and Pacific
biosciences SMRT technology to detect patterns in HIV integration and splicing
under various experimental conditions. Host-HIV interactions are reflected as
changes to these patterns when we either perturb specific host factors (Chapters 2
and 3), or investigate alternate cell types (Chapter 5). The data we generated
provides unique mechanistic insights into two steps of the HIV replication cycle.
Our observations generate fresh hypotheses about viral dependence on host factors,
and suggest new potential targets for antiretroviral therapeutics.
References


27


CHAPTER 2 – HIV INTEGRATION TARGETING: A PATHWAY

INVOLVING TRANSPORTIN-3 AND RAN-BINDING PROTEIN 2

Much of the contents of this chapter have been published as part of:
*equal contribution

Abstract

Genome-wide siRNA screens have identified host cell factors important for efficient HIV infection, among which are nuclear pore proteins such as RanBP2/Nup358 and the karyopherin Transportin-3/TNPO3. Analysis of the roles of these proteins in the HIV replication cycle suggested that correct trafficking through the pore might facilitate the subsequent integration step. Here we present data for coupling between these steps by demonstrating that depletion of Transportin-3 or RanBP2 altered the terminal step in early HIV replication, the selection of chromosomal sites for integration. We found that depletion of Transportin-3 and RanBP2 altered integration targeting for HIV. These knockdowns reduced HIV integration frequency in gene-dense regions and near gene-associated features, a pattern that differed from that reported for depletion of the HIV integrase binding cofactor Psip1/LEDGF/p75. MLV integration was not affected by the Transportin-3 knockdown. Using siRNA knockdowns and integration targeting analysis, we also implicated several additional nuclear proteins in proper target site

32
selection. To map viral determinants of integration targeting, we analyzed a chimeric HIV derivative containing MLV gag, and found that the gag replacement phenocopied the Transportin-3 and RanBP2 knockdowns. Thus, our data support a model in which Gag-dependent engagement of the proper transport and nuclear pore machinery mediates trafficking of HIV complexes to sites of integration.

**Introduction**

To complete the early steps of infection, retroviral preintegration complexes (PICs) must access the nucleus of the infected cell and integrate the viral cDNA into host chromatin. Gammaretroviruses such as MLV require nuclear envelope breakdown during mitosis to access cellular chromosomes and complete integration [1,2]. In contrast, lentiviruses such as HIV can enter the nucleus in non-cycling cells, presumably by traversing the nuclear pore [3-5].

Passage through the pore is likely a preferred route of nuclear entry for HIV-1 even in dividing cells – several components of the nuclear pore are required for efficient infection of dividing cells, even though PICs might also access the nucleus during nuclear breakdown in mitosis [6-11]. Moreover, in infections initiated during interphase, integration occurs before mitosis, while integration in cells infected just prior to mitosis is delayed until the following interphase [12]. These data suggest that the steps of HIV import through the nuclear pore may be coupled to subsequent integration. In support of this hypothesis, König and colleagues found that in dividing cells depleted of some nuclear pore factors or karyopherins,
HIV DNA entered the nucleus but did not integrate efficiently [9]. Thus the route of nuclear entry may influence subsequent integration, and the pore may provide the preferred route even in dividing cells.

Several host components are known to modulate retroviral integration. Integration target site selection is guided by the genomic environment of the integration acceptor site [13-18]. Lentiviruses such as HIV show a preference for integration in active transcription units, which may promote efficient expression after integration [17,19-21]. gammaretroviruses such as MLV show a preference for integration near gene 5’ ends and CpG islands [16-18]. Target site preferences of HIV integration are due in part to tethering by a host chromatin binding protein, LEDGF/p75 (product of the PSIP1 gene), which binds lentiviral IN [22,23] and mediates IN-chromatin binding [24,25]. In the absence of LEDGF/p75, HIV integration is severely compromised and integration in transcription units is diminished [26-28]. Recently, the tethering model for LEDGF/p75 function was bolstered by the finding that fusion proteins containing the IN-binding domain of LEDGF/p75 fused to alternative chromatin binding domains retargeted lentiviral integration efficiently [29-31].

Here we analyze host factors identified in genome-wide siRNA screens [6,9,11] and find links between transport into the nucleus and subsequent integration targeting. We chose factors whose depletion, like that of LEDGF/p75, led to an infection block at nuclear entry or integration. We initially surveyed effects of knocking down expression of ten genes, and then focused on two of them,
TNPO3 and RANBP2, which encode components of the nuclear pore and import machinery. TNPO3 encodes Transportin-3, a karyopherin [32] that has been shown to be required for import of HIV PICs into the nucleus in cycling cell lines and macrophages [6,7,9]. RanBP2 (originally named Nup358) is a large cyclophilin-related nuclear pore protein involved in the Ran-GTPase cycle that orchestrates much of nuclear import and export [33], and is also required for import of HIV PICs [9]. Recently, Lee and colleagues isolated a capsid mutant (N74D) [34] that bypassed the requirement for Transportin-3 and RanBP2, but acquired a requirement for other nuclear pore factors. HIV capsid had previously been suggested to be a viral determinant of nuclear entry [35] and these data suggest a possible direct interaction of capsid with Transportin-3 and RanBP2.

Using RNA interference, we reduced the expression of candidate genes, confirmed that HIV titer was reduced as a result, and then investigated the distribution of integration sites in the human genome using DNA bar coding and 454/Roche pyrosequencing. As controls, we studied infections and targeting by MLV. We also studied integration targeting by a derivative of HIV containing the gag gene (encoding the capsid structural proteins) of MLV. We found that depletion of Transportin-3 and RanBP2 resulted in marked alterations in the distribution of HIV integration sites, providing a link between nuclear entry and integration targeting. MLV integration patterns were not altered in Transportin-3 knockdowns, and substitution of MLV Gag into HIV phenocopied the effects of the knockdowns. Several additional host gene products were also identified as candidate members of
the pathway. Thus we can begin to specify a "railroad track" through the nuclear pore to favored sites of HIV DNA integration.

**Materials and Methods**

**Cell culture and viral infections**

For studies in HEK-293T cells VSV-G pseudotyped HIV vector particles were produced in HEK 293T cells by Lipofectamine (Invitrogen) transfection of p156RRLsin-PPTCMVGFPWPRE [36], the packaging construct pCMVdeltaR9 [37], and the vesicular stomatitis virus G-producing plasmid pMD.G. VSV-G pseudotyped MLV particles were produced in a similar manner but using the MLV vector segment (pMX-eGFP) and packaging construct pCGP (pCGP, gift of Paul Bates). For studies in primary macrophages, VSV-G pseudotyped HIV-Luc vector particles were produced as above using plasmids pNL4-3-R-E [9] and the vesicular stomatitis virus G-producing plasmid pMD.G. The replication-competent dual tropic clinical isolate HIV-1_{89,6} [38] was obtained from the University of Pennsylvania Center for AIDS Research (CFAR). The concentration of p24 in viral preparations was determined by ELISA by the University of Pennsylvania CFAR.

HEK 293T cells were grown in D10 media (DMEM supplemented with 10% FBS and 50μg/μL Gentamicin). For gene knockdowns, cells were grown to confluency, trypsinized and reverse transfected (100,000 cells/well in 12 well plates, 50,000/well in 24 well plates, and 8,000/well in 96 well plates) using RNAiMax (Invitrogen) with 25 pmol/mL siRNA. The siRNAs were purchased from
Qiagen (Qiagen) and are listed in Table 2-1. Transfection media was replaced after 48hr by 500μL of D10 plus 5ug DEAE dextran and virus in 12 well plates. Two viral inoculums were used (0.06μL or 1μL concentrated virus stock corresponding to 1.32ng or 22ng p24 per well, values determined by titration to result in infection of 30-60% or 80-100% of cells, respectively). Virus-containing media was replaced after 10-12 hours with 1mL D10 and incubated an additional 38 hours before harvest. Percent infection was measured using GFP fluorescence, which is not strongly affected by integration site placement in the HIV-based vectors with strong artificial promoters used here [39].

HIV infection and targeting rescue experiments in HEK293T cells were performed as described for siRNA knockdowns but with the co-transfection of siRNA-resistant or empty expression vectors (333ng plasmid/mL). The siRNA-resistant TNPO3 allele was constructed by introducing six conservative mutations in the third position of each codon and an N-terminal 3xFLAG-tag into the TNPO3 cDNA amplified HEK-293T cells. This product was then cloned into the mammalian expression vector pLNCX (kind gift of Paul Bates), engineered to contain a WPRE. Oligonucleotides used are listed in Table 2-2.

Primary monocytes were isolated from whole blood by Martha Kienzle in Ron Collman's lab by gel elutriation leading to 99% purity. Monocytes were incubated in 10 cm plates at a concentration of 1.2x10^6 cells/mL in RPMI supplemented with 10% FBS and 50 ng/mL MCSF (Peprotech) for 5-6 days to allow differentiation in to monocyte-derived-macrophages (MDM). One day prior to use,
<table>
<thead>
<tr>
<th>Target</th>
<th>siRNA Number</th>
<th>siRNA sequence</th>
<th>Product ID</th>
<th>Product Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>AllStarDeath</td>
<td>Proprietary</td>
<td>1027298</td>
<td>Hs_ANAPC2_1</td>
<td>AllStars Hs Cell Death Control siRNA</td>
</tr>
<tr>
<td>ANAPC</td>
<td>1</td>
<td>GAGAGTCTATATGCAAGTAA</td>
<td>Sio0101829</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>AAAAGTTTCTTACCCGACTA</td>
<td>Sio2652951</td>
<td>Hs_ANAPC2_5</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>CTCATGGATGGATCCTCAAA</td>
<td>Sio2653203</td>
<td>Hs_ANAPC2_6</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>ACGAGGCGGTGATTTGCTGTA</td>
<td>Sio4439456</td>
<td>Hs_ANAPC2_8</td>
</tr>
<tr>
<td>AQR</td>
<td>2</td>
<td>CAGGATGATATCGGGACTA</td>
<td>Sio0299530</td>
<td>Hs_AQR_2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>CAGGCCCTATTTGAAGTCAAT</td>
<td>Sio4168675</td>
<td>Hs_AQR_5</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>CCCTAGGGGTGCTCCAAATCAG</td>
<td>Sio421789</td>
<td>Hs_AQR_6</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>TGGCTGATCTGATGATGAACTA</td>
<td>Sio4368659</td>
<td>Hs_AQR_7</td>
</tr>
<tr>
<td>GL2 (irrel siRNA)</td>
<td>ACGGACGGGAGAATATCTCGA</td>
<td>1022070</td>
<td>Luciferase GL2 siRNA</td>
<td></td>
</tr>
<tr>
<td>IK</td>
<td>1</td>
<td>AACCCGTAACAAGAAGCTTGA</td>
<td>Sio0084336</td>
<td>Hs_IK_1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>CAGCCGATATGCAACAAAACA</td>
<td>Sio0084350</td>
<td>Hs_IK_3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>CAGGGCGCTCAAGGAAACCAA</td>
<td>Sio0084357</td>
<td>Hs_IK_4</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>AATGACATTGTAGTATTCAAG</td>
<td>Sio3000531</td>
<td>Hs_IK_9</td>
</tr>
<tr>
<td>MAP4</td>
<td>3</td>
<td>CCCAGATTTCAAATCTTATA</td>
<td>Sio0627809</td>
<td>Hs_MAP4_3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>AAGATCTTTATGGATTTGTA</td>
<td>Sio0627816</td>
<td>Hs_MAP4_4</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>CAGGGGAGATCAGCCTTAAA</td>
<td>Sio4315731</td>
<td>Hs_MAP4_8</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>AAGCGGAAACGCTGATTTAAA</td>
<td>Sio4337053</td>
<td>Hs_MAP4_9</td>
</tr>
<tr>
<td>NUP98</td>
<td>3</td>
<td>CTGGAGTTAGACTAACATTAA</td>
<td>Sio1115332</td>
<td>Hs_NUP98_3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>CAGTGATATTGTGACATGAA</td>
<td>Sio2635115</td>
<td>Hs_NUP98_5</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>AACCTATTGCAAACCTTATT</td>
<td>Sio3028578</td>
<td>Hs_NUP98_7</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>CTCACTAAAGTTGTTGACTTAA</td>
<td>Sio3088491</td>
<td>Hs_NUP98_8</td>
</tr>
<tr>
<td>PRPF38A</td>
<td>1</td>
<td>TCCAGCGGAATAAGATGGA</td>
<td>Sio0395801</td>
<td>Hs_PRPF38A_1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>CAGCCGCTATGCTGAGGATA</td>
<td>Sio4159225</td>
<td>Hs_PRPF38A_2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>TCCACCGGACTCGATAATCAA</td>
<td>Sio4186567</td>
<td>Hs_PRPF38A_3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>CCAGCTCTGAGGTCGCTTAAA</td>
<td>Sio4246634</td>
<td>Hs_PRPF38A_4</td>
</tr>
<tr>
<td>RanBP</td>
<td>2</td>
<td>CGGTTGGGCAAGCAGAGTTAAC</td>
<td>Sio0698650</td>
<td>Hs_RANBP2_2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>AGGCCGGTGGTTGAGAATATAA</td>
<td>Sio3043901</td>
<td>Hs_RANBP2_5</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>TCCTTGCGATATGCAATCCCA</td>
<td>Sio3117310</td>
<td>Hs_RANBP2_6</td>
</tr>
<tr>
<td>SNW1</td>
<td>2</td>
<td>TCGGGATGAGAATAATGCGAAA</td>
<td>Sio0098609</td>
<td>Hs_SKIIP_2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>CACAGCGGCTTCCAAAGCTGAA</td>
<td>Sio0098616</td>
<td>Hs_SKIIP_3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>AACGCTTCGGCTGCTGAGGAA</td>
<td>Sio3001812</td>
<td>Hs_SKIIP_5</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>CCAGTAGAAGACGCTTTAAA</td>
<td>Sio2655478</td>
<td>Hs_SKIIP_7</td>
</tr>
<tr>
<td>TNPO3</td>
<td>1</td>
<td>ACCGAACTGCTTATGGAATCA</td>
<td>Sio1000513</td>
<td>Hs_TNPO3_1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>CTGGAGATCCCTAGTGTTTA</td>
<td>Sio1000520</td>
<td>Hs_TNPO3_2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>CAAAGTGTCATGACATTTAATC</td>
<td>Sio1000534</td>
<td>Hs_TNPO3_4</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>CTGGGAGATCAGGTGTTGGA</td>
<td>Sio2639721</td>
<td>Hs_TNPO3_5</td>
</tr>
<tr>
<td>WDHD1</td>
<td>1</td>
<td>CTGGACTTTGAGTGAACACTAAA</td>
<td>Sio0094472</td>
<td>Hs_WDHD1_1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>CAGGAGATGAAATGCTTCTCCTTTTA</td>
<td>Sio0094479</td>
<td>Hs_WDHD1_2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>TGCTGGTGAATTTGACCTTATA</td>
<td>Sio0094493</td>
<td>Hs_WDHD1_4</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>CAGGAGATTATATGCTCTGAA</td>
<td>Sio2638503</td>
<td>Hs_WDHD1_5</td>
</tr>
<tr>
<td>WDR46</td>
<td>1</td>
<td>CAGCGGCGCCTCCTGTAACCAAGA</td>
<td>Sio3071383</td>
<td>Hs_WD46_1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>CAGTGGGCGCTTACCTAGACTA</td>
<td>Sio3073609</td>
<td>Hs_WD46_2</td>
</tr>
<tr>
<td></td>
<td>11-2</td>
<td>TGCCCTCGTATGCGTAAAACCA</td>
<td>Sio0077455</td>
<td>Hs_C6orf11_2</td>
</tr>
<tr>
<td></td>
<td>11-4</td>
<td>AAAGAAGTTCGCGATCTTCAA</td>
<td>Sio0077469</td>
<td>Hs_C6orf11_4</td>
</tr>
</tbody>
</table>
the non-adherent cells were resuspended in fresh media with MCSF and restored to the adherent cells.

Adherent MDMs were removed from plates with rubber spatula and were plated in 96 well plates at 25x10^3 cells per well and reverse-transfected with siRNAs as above in the presence of 12.5 ng/mL MCSF. For infections, a volume of HIV-Luc equivalent to 150 ng p24 or of HIV-1.89.6 equivalent to 60 ng p24 was added to cells in 50 uL media total. Cells were spinoculated 1hr at room temperature and then incubated 4-5 d at 37°C in the presence or absence of Raltegravir obtained through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH: Raltegravir (Cat # 11680) from Merck & Company, Inc..

Infection by HIV-Luc was assayed as luciferase production in triplicate in three independent experiments using 48 uL of the Promega BrightGlo reagent per well according to the manufacturer’s protocol.

**Immunofluorescence Microscopy**

MDMs were assayed for infection by HIV-1.89.6 at 3dpi by immunofluorescence microscopy. MDMs were fixed in 4% formaldehyde, blocked in 2% BSA in 0.1% TritonX in PBS, and incubated with Rabbit-anti-HIV p24 (NIH AIDS reagent #4250) diluted 1:5000 in 0.1% TritonX in PBS overnight at 4°C. MDMs were washed and incubated 1hr at room temperature in secondary Alexa-fluor 594 conjugated goat-anti-rabbit antibody diluted 1:2500 in 0.1% TritonX in PBS containing 1:3000 diluted DAPI stain (kind gift of Jianxin You). Microscopy was
performed on the inverted fluorescence microscope IX81 (Olympus) equipped with UPlanSApo 40/0.95 NA lens (Olympus), UPlanSApo 100/1.4 oil immersion lens (Olympus), and a high-resolution charge-coupled device camera (QImaging, FAST1394), using SlideBook 5.0 imaging software from Intelligent Imaging Innovations, Inc. Infected MDMs stained only with the secondary antibody showed no background fluorescence.

Toxicity Assays

Toxicity of siRNAs was measured 48 hr after transfection both visually and by the CellTiter-Glo Luminescent Cell Viability Assay (Promega) following the manufacturer’s instructions. Cells were reverse transfected in 96 well plates with the indicated siRNAs at 25 pmol/ml final concentration and incubated at 37°C.

Quantitative PCR

RNA was purified from cells harvested 48 hr post siRNA transfection using either the RNeasy Mini Kit (Qiagen) or the RNAspin Mini Kit (GE Healthcare) per manufacturer’s instructions. RT-PCR was carried out using the High Capacity RNA to cDNA Kit (Applied Biosystems) and relative RNA levels were measured by the ddCt method using Taqman Gene Expression Assays (Applied Biosystems) with GUSB as the internal reference according to manufacturer’s instructions using the Applied Biosystems 7500 Fast Real Time PCR system. Assays IDs were Hs00193785_m1, Hs00600887_m1, Hs00173172_m1, Hs00273527_m1,
Hs00159048_m1, Hs00610583_m1, Hs01108576_m1, Hs00203499_m1, 
Hs00273351_m1, Hs00180522_m1 for genes measured for knockdown and product 
number 4333767F for the GUSB endogenous control assay.

**Immunoblotting**

Cells harvested 48 hr post siRNA transfection were lysed in 150 mM NaCl, 50 
mM tris-HCl pH 8.0, 2 mM EDTA, 1% NP-40, 0.5% deoxycholic acid, 0.1% SDS with 
protease inhibitors (Roche, 1697498). Protein was separated by SDS-PAGE using 4- 
15% Tris-HCl gradient gels (Biorad). Protein levels were measured by 
immunoblotting using antibodies against Transportin-3 (ab54353, Abcam) and 
RanBP2 (ab2938, Abcam). HRP conjugated secondary antibodies (p0260, DAKO 
A/S, Denmark, and ab6721-1, Abcam) were used for detection with SuperSignal 
West Pico Chemiluminescent Substrate (Thermo Scientific, Pierce Protein Research 
Products). Beta-tubulin was used as a loading control, detected by the HRP 
conjugated antibody (ab21058, Abcam).

**Integration site amplification**

Purified genomic DNA was digested overnight with Msel, ligated at 16°C to 
PCR adapters, and digested a second time with Sacl. Barcoded nested PCR was then 
performed using primers and conditions described previously [39,40]. 
Amplification products between 200-600bp were gel-excised, purified, and 
sequenced on a Genome Sequencer FLX Titanium Series (Roche 454 Sequencing) at
the University of Pennsylvania or the University of Florida. DNA oligonucleotides used in this study are listed in Table 2-2.

<table>
<thead>
<tr>
<th>Oligo</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mse linker+</td>
<td>GTAATACGACTCAGCTATAGGGCTCCGCTTAAGGGAC</td>
<td>Linker oligo plus strand</td>
</tr>
<tr>
<td>Mse linker-</td>
<td>[Phosp]TAGTCCCTTAAGCGGAG-[AmC7-Q]</td>
<td>Linker oligo minus strand, 5'phosphate and 3' amide C7 modification</td>
</tr>
<tr>
<td>MseL1</td>
<td>GTAATACGACTCAGCTATAGGGC</td>
<td>Linker primer for first round PCR</td>
</tr>
<tr>
<td>MKL-3</td>
<td>CTTCAGCTCAATAAGCTTGCTTGA</td>
<td>HIV LTR primer for first round PCR</td>
</tr>
<tr>
<td>GW-3</td>
<td>GCCTCCCTCGCCACATCGAGGGCTCCGCTTAAGGGAC</td>
<td>Linker primer for second round PCR</td>
</tr>
<tr>
<td>B-HIV</td>
<td>GCCTTGCAGCCGCTCGAGXXXXXXXAGACCCTTTAGTCAGTTGGAATTC</td>
<td>HIV LTR primer for second round PCR, Xs: barcode for sequencing</td>
</tr>
<tr>
<td>KR-5</td>
<td>CCTCTGAGCTGATGTCAGC</td>
<td>MLV LTR primer for first round PCR</td>
</tr>
<tr>
<td>B-MLV</td>
<td>GCCTTGCAGCCGCTCGAGXXXXXXXGATTGAGCTACCCGTACGGG</td>
<td>MLV LTR primer for second round PCR, Xs: barcode for sequencing</td>
</tr>
<tr>
<td>kr237</td>
<td>GTACCTTTGCTGCGGAGCCAGCGTCGATGGAGTTTCCACATTTTGCAGTTGGAATTC</td>
<td>TNPO3 quikchange fwd siRNA_4</td>
</tr>
<tr>
<td>kr238</td>
<td>GTAGAGGCAATGGCCCATAATGGGAATCAGCAGCTGGCTCCGCAAGCAAAGGGGAGC</td>
<td>TNPO3 quikchange rev siRNA_4</td>
</tr>
</tbody>
</table>

**Analysis**

Integration datasets used in this study are listed in Table 2-3. Sequences that began within three base pairs of the LTR end and showed unique best alignments to the human genome by BLAT (hg18, version 36.1, >98% match score) were considered true integration sites and used in the below analysis. Identical
integration sites identified in two or more separately amplified samples were considered to be PCR contamination and were omitted.

Comparisons to genomic features were carried out as described previously [13] using a combination of conditional logit regression and Bayesian model averaging. For each integration site three matched random control sites (MRCs) were computationally generated that were the same distance from an Msel restriction site as the corresponding experimental site but otherwise randomly distributed in the human genome. These sites were used to control for biases in isolating integration sites due to uneven distributions of the Msel restriction site. Details of statistical methods are available in [13,39-43]. Methods used to determine associations with multiple genomic features are summarized in Appendix 1 of this dissertation. Gene expression analyses utilized data from 293T cells [27] with expression measured using the Affymetrix HU133 plus 2.0 gene chip array. All integration site sequences have been deposited in the NCBI dbGSS database with accession number LIBGSS_038750.

Entrez Gene ID Numbers for Genes in this study

Table 2-3 - Integration Site Datasets used in this study.

<table>
<thead>
<tr>
<th>siRNA Target*</th>
<th>siRNA Number</th>
<th>Virus**</th>
<th>Target Cell Type</th>
<th># Unique integration sites</th>
<th>References</th>
<th>Generated By</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mock</td>
<td></td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>270</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>Mock</td>
<td></td>
<td>HIV</td>
<td>293T</td>
<td>43</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>Luciferase</td>
<td></td>
<td>HIV</td>
<td>293T</td>
<td>4296</td>
<td>this work</td>
<td>Keshet Ronen</td>
</tr>
<tr>
<td>ANAPC2</td>
<td>5</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>295</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>ANAPC2</td>
<td>5</td>
<td>HIV</td>
<td>293T</td>
<td>316</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>ANAPC2</td>
<td>6</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>120</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>ANAPC2</td>
<td>8</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>197</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>ANAPC2</td>
<td>8</td>
<td>HIV</td>
<td>293T</td>
<td>107</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>IK</td>
<td>9</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>641</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>IK</td>
<td>9</td>
<td>HIV</td>
<td>293T</td>
<td>376</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>LEDGF/p75</td>
<td>shRNA</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>462</td>
<td>this work, cells: Ciuffi et al. [26]</td>
<td>Keshet Ronen</td>
</tr>
<tr>
<td>MAP4</td>
<td>4</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>444</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>MAP4</td>
<td>4</td>
<td>HIV</td>
<td>293T</td>
<td>38</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>MAP4</td>
<td>9</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>419</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>MAP4</td>
<td>9</td>
<td>HIV</td>
<td>293T</td>
<td>641</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>MAP4</td>
<td>3</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>3696</td>
<td>this work</td>
<td>Keshet Ronen</td>
</tr>
<tr>
<td>MAP4</td>
<td>4</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>3394</td>
<td>this work</td>
<td>Keshet Ronen</td>
</tr>
<tr>
<td>NUP98</td>
<td>3</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>128</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>NUP98</td>
<td>3</td>
<td>HIV</td>
<td>293T</td>
<td>94</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>NUP98</td>
<td>5</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>171</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>NUP98</td>
<td>5</td>
<td>HIV</td>
<td>293T</td>
<td>493</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>NUP98</td>
<td>7</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>125</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>NUP98</td>
<td>7</td>
<td>HIV</td>
<td>293T</td>
<td>75</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>NUP98</td>
<td>8</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>84</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>NUP98</td>
<td>8</td>
<td>HIV</td>
<td>293T</td>
<td>88</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>RANBP2</td>
<td>6</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>458</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>RANBP2</td>
<td>6</td>
<td>HIV</td>
<td>293T</td>
<td>344</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>RANBP2</td>
<td>pool</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>470</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>RANBP2</td>
<td>pool</td>
<td>HIV</td>
<td>293T</td>
<td>363</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>SNW1</td>
<td>5</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>153</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>SNW1</td>
<td>5</td>
<td>HIV</td>
<td>293T</td>
<td>527</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>TNPO3</td>
<td>1</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>75</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>TNPO3</td>
<td>1</td>
<td>HIV</td>
<td>293T</td>
<td>51</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>TNPO3</td>
<td>2</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>165</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>TNPO3</td>
<td>2</td>
<td>HIV</td>
<td>293T</td>
<td>510</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>TNPO3</td>
<td>4</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>171</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>TNPO3</td>
<td>4</td>
<td>HIV</td>
<td>293T</td>
<td>260</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>TNPO3</td>
<td>5</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>600</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>TNPO3</td>
<td>5</td>
<td>HIV</td>
<td>293T</td>
<td>297</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>TNPO3</td>
<td>pool</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>186</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>TNPO3</td>
<td>pool</td>
<td>HIV</td>
<td>293T</td>
<td>216</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>WDHD1</td>
<td>4</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>240</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>siRNA Target*</td>
<td>siRNA Number</td>
<td>Virus**</td>
<td>Target Cell Type</td>
<td># Unique integration sites</td>
<td>References</td>
<td>Generated By</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>---------</td>
<td>-----------------</td>
<td>---------------------------</td>
<td>------------</td>
<td>--------------</td>
</tr>
<tr>
<td>WDHD1</td>
<td>4</td>
<td>HIV</td>
<td>293T</td>
<td>7</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>WDR46</td>
<td>1</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>178</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>WDR46</td>
<td>1</td>
<td>HIV</td>
<td>293T</td>
<td>220</td>
<td>this work</td>
<td>Troy Brady</td>
</tr>
<tr>
<td>WDR46 pool (1,2,C6orf11_2, C6orf11_4)</td>
<td>HIV/high MOI</td>
<td>293T</td>
<td>260</td>
<td>this work</td>
<td>Troy Brady</td>
<td></td>
</tr>
<tr>
<td>WDR46 pool (1,2,C6orf11_2, C6orf11_4)</td>
<td>HIV</td>
<td>293T</td>
<td>292</td>
<td>this work</td>
<td>Troy Brady</td>
<td></td>
</tr>
<tr>
<td>PRPF38A</td>
<td>2</td>
<td>HIV</td>
<td>293T</td>
<td>1577</td>
<td>this work</td>
<td>Keshet Ronen</td>
</tr>
<tr>
<td>PRPF38A</td>
<td>3</td>
<td>HIV</td>
<td>293T</td>
<td>705</td>
<td>this work</td>
<td>Keshet Ronen</td>
</tr>
<tr>
<td>GL2+EV</td>
<td></td>
<td>HIV</td>
<td>293T</td>
<td>1226</td>
<td>this work</td>
<td>Karen Ocwieja</td>
</tr>
<tr>
<td>TNPO3+R</td>
<td>4</td>
<td>HIV</td>
<td>293T</td>
<td>1492</td>
<td>this work</td>
<td>Karen Ocwieja</td>
</tr>
<tr>
<td>TNPO3+EV</td>
<td>4</td>
<td>HIV</td>
<td>293T</td>
<td>314</td>
<td>this work</td>
<td>Karen Ocwieja</td>
</tr>
<tr>
<td>GL2+EV</td>
<td></td>
<td>MLV</td>
<td>293T</td>
<td>616</td>
<td>this work</td>
<td>Karen Ocwieja</td>
</tr>
<tr>
<td>TNPO3+R</td>
<td>4</td>
<td>MLV</td>
<td>293T</td>
<td>354</td>
<td>this work</td>
<td>Karen Ocwieja</td>
</tr>
<tr>
<td>TNPO3+EV</td>
<td>4</td>
<td>MLV</td>
<td>293T</td>
<td>376</td>
<td>this work</td>
<td>Karen Ocwieja</td>
</tr>
<tr>
<td>SNW1+EV</td>
<td>5</td>
<td>MLV</td>
<td>293T</td>
<td>90</td>
<td>this work</td>
<td>Karen Ocwieja</td>
</tr>
<tr>
<td>ANAPC2+EV</td>
<td>5</td>
<td>MLV</td>
<td>293T</td>
<td>79</td>
<td>this work</td>
<td>Karen Ocwieja</td>
</tr>
<tr>
<td>ANAPC2+EV</td>
<td>6</td>
<td>MLV</td>
<td>293T</td>
<td>32</td>
<td>this work</td>
<td>Karen Ocwieja</td>
</tr>
<tr>
<td>ANAPC2+EV</td>
<td>8</td>
<td>MLV</td>
<td>293T</td>
<td>251</td>
<td>this work</td>
<td>Karen Ocwieja</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td>HIVPuro</td>
<td></td>
<td></td>
<td>485</td>
<td>Lewinski et al. [49]</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td>HIVmGag</td>
<td>HeLa</td>
<td></td>
<td>452</td>
<td>Lewinski et al. [49]</td>
</tr>
</tbody>
</table>

* GL2 luciferase targeted; EV, siRNA transfected with empty vector plasmid; R, siRNA transfected with siRNA insensitive rescue plasmid.

**Infections performed to infect 30-60% of untreated cells except "highMOI" where 90-100% were infected

---

**Results**

**Surveying integration site distributions after siRNA knockdown**

This study was conducted in collaboration with Keshet Ronen and Troy Brady from the Bushman Lab. Keshet and Troy initially analyzed 10 genes previously implicated as HIV cofactors at or near the integration step to determine whether they had effects on integration targeting. They selected NUP98 [8,9], MAP4 [6,9], IK [9], ANAPC2 [9,11], PRPF38A [9], RANBP2 [6,9], SNW1 [9], and TNPO3
[6,9] from siRNA screens, and two other genes, WDR46 and WDHD1, the products of which bind LEDGF/p75 in yeast two-hybrid screens (unpublished data). For each gene, they tested several different siRNAs in HEK-293T cells. With the exception of WDR46, for each gene, at least one siRNA was chosen that conferred ≥40% inhibition of infection by a VSVG-pseudotyped GFP reporter virus, as defined as the percent of cells expressing the GFP marker 48h after infection. Collectively we assessed toxicity of the siRNAs (Fig. 2-1) and confirmed reduction of mRNA levels by quantitative RT-PCR (Fig. 2-2A). Selected knockdowns were verified by Western blot (Fig. 2-2B). We also confirmed that TNPO3 knockdown blocked infection of primary monocyte-derived-macrophages (Fig. 2-3).

The initial scan by Keshet and Troy showed robust effects on infection efficiency in HEK293T cells for the nuclear import factors Transportin-3 and RanBP2, confirming observations from earlier studies [6,7,9,44]; therefore, these genes were studied in detail as described in the following sections. Results for Transportin-3 and RanBP2 have been corroborated by further studies using stable knockdowns with shRNAs in HeLa cells that achieved efficient reductions in mRNA levels (discussed in Chapter 3 and [45]). Keshet and Troy also analyzed the remaining 8 genes for integration targeting using our high throughput pipeline. We return to findings for this group of genes at the end of the Results.
**Figure 2-1 - Cell viability after siRNA transfection.** Toxicity of siRNAs was measured as a decrease in cell viability represented by the amount of ATP released from lysed cells 48hr after transfection, and assayed as light released from the ATP-dependent luciferase reaction. An siRNA cocktail known to deplete anti-apoptotic factors, siDeath, was used as a positive control for toxicity. GL2, not present in human cells, was targeted with siRNA as a negative control. All values normalized to GL2 controls. Data shown is representative of at least two independent experiments.
Figure 2-2 - siRNA Depletion of Targeted factors. (A) Quantitative PCR showed decreased expression of the targeted factors relative to expression of housekeeping gene GUSB. All values were normalized to those in cells treated with the control siRNA against GL2. Error bars indicate standard error of four measurements. Data presented is representative of at least three replicate experiments. (B) Reduction in Transportin-3, RanBP2, and SNW1 protein levels after RNAi. Protein abundance was measured at the time of infection by Western blot with β-tubulin as a loading control. For comparison, protein levels are shown in cells treated with an siRNA against GL2.
Figure 2-3 - Transportin-3 is required for infection of macrophages. Primary monocytes were isolated from healthy human donors and differentiated in culture. (A) TNPO3 expression was confirmed to be depleted relative to GUSB expression in siRNA treated cells by quantitative PCR. Data is normalized to expression in siGL2 treated cells. Error bars indicate standard error of four measurements. Decreased infection of Transportin-3-depleted monocyte derived macrophages was observed using a VSV-G pseudotyped luciferase (GL3) reporter virus (B) and by immunofluorescence microscopy of cells infected with a syncitia-inducing clinical isolate HIV-1 \textsubscript{1996} (C). Blue, DAPI stain, Red, HIV p24 (capsid). Raltegravir was used as a control for inhibition of infection. Error bars in (B) indicate standard deviation of triplicate measurements in three separate experiments.

HIV integration site selection is modified by depletion of Transportin-3 and RanBP2

Having confirmed that knockdown of Transportin-3 and RanBP2 reduced the efficiency of HIV infection in HEK-293T cells, Troy and Keshet examined the effect of these factors on integration site selection. Troy used ligation-mediated PCR and 454-pyrosequencing to identify integration sites as previously described [39,43].
Recovered genomic sequences were mapped to the human genome draft hg18. Association of integration sites with genomic features was then assessed.

In the human genome, many types of features are linked – for example, gene dense regions are rich in CpG islands and DNaseI sites, high in G/C content, and rich in highly expressed genes [46,47]. As a first step in illustrating the results, we will discuss integration site distributions as a function of gene density. In cells depleted of Transportin-3 or RanBP2, the distribution of HIV integration sites was altered towards regions of lower gene density in comparison to control cells treated with siGL2, which targets firefly luciferase GL2, a gene not found in the HEK-293T cells (Fig. 2-4A). The trend towards integration in less gene dense regions was significant for both RANBP2 and TNPO3 knockdowns (p< 0.001). There was no evidence of a bimodal distribution of integration sites with respect to gene density in the knockdown cells, which might have reflected knockdown of the factors in only a portion of the cells. However, HIV integration sites in control cells did show a subtle bimodal distribution with respect to gene density. In the knockdown cells the peak of integration sites in more gene dense regions is disproportionately decreased, suggesting that Transportin-3 and RanBP2 direct integration targeting of a subset of integration events (see Discussion).

The average gene density at integration sites in cells depleted of RANBP2 and TNPO3, each by multiple different siRNAs, was reduced compared to cells treated with siGL2, though it remained higher than the gene density at matched random control sites (Fig. 2-4B). Thus integration in gene dense regions is promoted in part
by RanBP2 and Transportin-3. As a control for the fact that the knockdowns diminished infection, Keshet and Troy investigated whether infections at low MOI altered the distribution of integration sites, but MOI was not found to affect integration targeting detectably (data not shown).

Keshet and Troy analyzed integration frequency relative to a large collection of genomic features and observed a common set of changes in both the Transportin-3 and RanBP2 depleted cells relative to the controls. The reduction in integration in gene dense regions was significant for both TNPO3 and RANBP2 knockdowns when analyzed over multiple genomic intervals of different lengths. Significant differences were also seen when only expressed genes (identified by Affymetrix chip transcriptional profiling) were considered in a similar analysis. Genomic features that correlate with gene density such as DNase I hypersensitive sites and CpG islands were similarly enriched near control HIV integration sites but less enriched near sites from TNPO3 and RANBP2 knockdown cells. GC-rich regions, normally favored by HIV [17], were generally disfavored in the Transportin-3 and RanBP2 knockdowns.

By contrast, gene density at integration sites was not significantly affected in LEDGF/p75 knockdowns compared to the control. The GC content and the density of CpG islands within one kb of integration sites actually increased in LEDGF/p75-depleted cells [26-28], indicating divergent effects on integration targeting. Integration within genes, which is reproducibly diminished in LEDGF/p75-depleted cells [26-28], was not affected by TNPO3 knockdown, and showed only a slight
decrease in the RANBP2 knockdown cells. Together these data suggest that Transportin-3 and RanBP2 influence HIV integration targeting relative to a collection of features associated with gene dense regions, and do so in a manner that differs from LEDGF/p75 tethering.

**Figure 2-4 - Effects of siRNA treatments on HIV integration in gene dense regions.** Integration datasets for this analysis were generated by Keshet Ronen and Troy Brady. Cells were transfected with individual siRNAs or an siRNA pool of four siRNAs targeting the same gene as indicated, and infected 48 hr later for an additional period of 48 hr prior to integration site analysis. The number of genes in 1Mb windows surrounding each integration site or each computationally-generated matched random control (MRC) site was tabulated. (A) Histogram indicating distribution of integration sites with respect to gene density. Integration sites in each dataset were binned (along the X-axis) according to the number of genes within 1MB interval surrounding each site. Curves were computed from histogram plot using Gaussian kernal density estimates. (B) Barplot of the average number of RefSeq genes in 1 Mb windows surrounding sites of HIV integration or computationally generated MRCs. Mock transfected cells (no RNAi) and cells treated with the siRNA targeting luciferase GL2 (siGL2) are shown as controls. Asterisks denote significant difference from siGL2 treated cells as determined by the nonparametric Mann–Whitney test (*P < 0.05; **P < 0.01; ***P < 0.001).
Effect of Transportin-3 depletion on integration site selection can be partially rescued by expression from an siRNA insensitive TNPO3 allele

Keshet and Troy showed that multiple different siRNAs directed against TNPO3 and RANBP2 mRNAs yielded similar effects on integration targeting that were not observed in control knockdowns. Thus, off-target effects were unlikely to explain the observed alterations in integration targeting.

However, as an additional control, we analyzed complementation of the Transportin-3 depletion using a plasmid-encoded siRNA-insensitive allele generated by site-directed mutagenesis of the siRNA target sequence. The plasmid was created in collaboration with Alyssa Huegel, a graduate student at the University of Pennsylvania. The RANBP2 coding region is very large (11,711 bp), and so rescue experiments were not attempted for this factor. Co-transfection of the resistant Transportin-3 expression vector with the corresponding siRNA resulted in overexpression of Transportin-3 and restored HIV infection, increasing reporter virus GFP expression above control levels (Fig. 2-5A and B).

We observed an increase in gene density near integration sites in knockdown cells co-transfected with the siRNA-insensitive TNPO3 allele compared to vector-only controls (Fig. 2-5C and D). The average number of genes within 1Mb of HIV integration sites increased from 11 (in the presence of TNPO3 si4 and an empty vector) to 14 when Transportin-3 expression was rescued (p < 0.01).
Figure 2-5 - Transfection of a Transportin-3 allele insensitive to TNPO3 si4 restores protein expression, HIV infectivity, and partially restores wild-type HIV integration site distributions. (A) Western blot showing Transportin-3 is reduced after co-transfection with siRNA and empty vector, and overexpressed after co-transfection with siRNA and rescue plasmid. Endogenous levels of Transportin-3 are shown in cells transfected with the control siRNA targeting GL2 and an empty vector. (B) HIV infection in cells treated with TNPO3 si4 in the presence or absence of the Transportin-3 rescue plasmid. 48 hr after transfection cells were infected with a VSVG-pseudotyped HIV-1 vector carrying a GFP reporter. At 48 hpi cells were harvested and the percent of cells expressing GFP was determined by flow cytometry. The Y-axis shows relative infection compared to infection in the control (GL2 siRNA + empty vector-transfected) cells. (C) Average gene density in 1Mb windows surrounding HIV integration sites in cells depleted or rescued for Transportin-3 expression. Asterisks denote significant differences as determined by the Mann–Whitney test (*P < 0.05; **P < 0.01; ***P < 0.001). (D) Histogram indicating distribution of integration sites with respect to gene density as in Figure 2-4A.
The effect of knockdown on the presence and absence of rescue on additional genomic features is shown in Figure 2-6 (analysis described in Appendix 1). Depletion of TNPO3 in the presence of an empty vector resulted in a significantly decreased HIV integration preference for several features associated with gene density as was previously observed for both Transportin-3 and RanBP2 knockdowns (above). In siRNA treated cells in which TNPO3 was expressed from the siRNA resistant plasmid, much of the wild type pattern of integration sites was restored; however, features associated with gene density remained significantly less favored for integration than in wild type cells. It is unclear why restoring Transportin-3 protein levels did not fully rescue the integration defect, but this result may be due to the abnormally high levels of Transportin-3 expressed from the siRNA-resistant construct. Nevertheless, these data support the idea that off-target effects of the TNPO3 siRNA do not account for the phenotypes observed.

**Transportin-3 depletion has no detectable effect on gene density surrounding MLV integration sites.**

As a control, we tested whether MLV integration, which requires cell division for infection and is not dependent on Transportin-3 [7,48], showed altered integration targeting in the Transportin-3-depleted cells. We found that treatment with siRNA targeting TNPO3 mRNA, either in the presence or absence of the rescue plasmid, did not affect MLV infection efficiency (Fig. 2-7A). We sequenced MLV integration sites from knockdown and control cells and found no significant changes
**Figure 2-6 – Effects of Transportin-3 depletion on integration near multipie chromosomal features.** Genes targeted by siRNA in infected cells including the control, GL2, are shown above the columns. The genomic features analyzed are shown in the rows and labeled on the left. Relationships between integration frequency and feature density are summarized using ROC curve areas, where increasing shades of blue indicate a negative correlation with integration frequency and increasing shades of red indicate a positive correlation with integration frequency relative to matched random control distributions. The control GL2 siRNA set was used for pairwise statistical comparisons (overlay dashes). P values summarizing the significance of the departure from the GL2 control are shown with asterisks (*P < 0.05; **P < 0.01; ***P < 0.001). Note that the asterisks and the heat map summarize different comparisons (to siGL2 and matched random controls, respectively). The base pair values in the row labels indicate the size of the genomic interval used for analysis – often the most appropriate interval is not known, so several different interval sizes are compared. A detailed guide to the analysis in this figure can be found in Appendix 1.
in MLV integration frequency in gene dense regions (Fig. 2-7B), within transcription units, or with respect to GC content (data not shown). These data indicate that Transportin-3 depletion does not affect MLV integration targeting as it does for HIV.

Other nuclear factors may participate in directing integration to gene dense regions

Keshet and Troy also acquired integration site data sets for 293T cells treated with siRNAs for NUP98, MAP4, IK, ANAPC2, PRPF38A, SNW1, WDR46 and WDHD1, as well as for a previously characterized cell line stably depleted of LEDGF
[26]. For many of these, considerable toxicity was detected (Fig. 2-1). Thus interpretation of integration targeting results for these factors is more tentative than for Transportin-3 and RanBP2. Knockdown of several of the factors (ANAPC2, SNW1, PRPF38, WD1, and IK) led to decreased integration in gene dense regions. MAP4 depletion was also seen to modestly decrease integration preference for gene dense regions in some experiments. By contrast, gene density at integration sites in cells stably depleted of LEDGF was not significantly decreased compared to the siGL2 control. For two of these genes, SNW1 and ANAPC2, we confirmed that although MLV infection is diminished in the knockdowns as previously noted [9], the gene density at MLV integration sites is unchanged (Fig. 2-8), suggesting that, like Transportin-3, the factors encoded by these genes are potentially involved in targeting pathways specific for HIV.

![Graph](image1.png)

**Figure 2-8 - MLV infection and integration site distributions after siRNA treatment targeting SNW1 and ANAPC2.** (A) Infection levels of MLV in cells treated with the control siRNA targeting GL2 or siRNAs targeting human genes SNW1 and ANAPC2. Infection levels measured by flow cytometry as the percentage of cells expressing GFP encoded by the MLV vector (B) Average gene density in 1Mb windows surrounding MLV integration sites in cells depleted for SNW1 or ANAPC2.
For those knockdowns where at least 200 integration sites were sequenced (Table 2-3), the global integration site patterns were investigated by assessing integration frequency relative to many genomic features for each knockdown, and the patterns were clustered using a conditional logit model to conduct pairwise comparisons of the datasets. This analysis was performed by Chuck Berry, Ph.D.. The controls (Mock transfection and siGL2-transfected) clustered in a group separate from Transportin-3 and RanBP2 knockdowns. Data sets for several additional gene knockdowns clustered in the TNPO3/RANBP2 group, including IK, ANAPC2, SNW1, WDHD1 and PRPF38A. For MAP4 and WDR46 different siRNAs fell in different clusters, and so these have an indeterminate effect. Thus the IK, ANAPC2, SNW1, WDHD1 and PRPF38A genes encode candidates for additional factors acting in the same pathway with Transportin-3 and RanBP2. The LEDGF/p75 knockdown was an outlier in the control cluster. This is consistent with LEDGF/p75 knockdown leading to effects not seen in depletion of RanBP2, or Transportin-3. For this analysis both low MOI (30-60% infected wild type cells) and high MOI (90-100% infected cells) infections were investigated. In most cases the MOI made no difference on the overall position of a knockdown among the clusters, suggesting that the roles of the factors are not saturable under the conditions tested.

**HIV gag is a determinant of integration targeting to gene dense regions**

Members of the Bushman lab previously studied integration targeting in HeLa cells using HIV chimeras containing MLV gag, MLV IN, or both, in place of their
HIV counterparts [49]. MLV IN proved to be a dominant determinant of MLV-like integration, resulting in integration near transcription start sites by HIV derivatives containing MLV IN. Similar chimeric viruses have been used to show that HIV capsid is a dominant viral determinant of HIV nuclear entry in non-dividing cells [50]. At the time this work was in process, Lee and colleagues had recently suggested that the HIV CA protein might determine the interactions between HIV PICs and nuclear pore components [34]. These findings led us to reinvestigate integration targeting by the HIV chimera containing MLV gag in place of HIV gag (HIVmGag; Fig. 2-9A) [49]. Using previously generated integration site datasets [49], we found that HIVmGag showed a shift in distribution of integration sites towards less gene dense regions compared to the unmodified control (Fig. 2-9B). The average number of genes within 1MB of HIVmGag integration sites was 11 as compared to 20 for the unmodified HIV control (A Chi square test over ranked comparisons of gene density values between the two sets attains a p value <2.22^-16). A comparison over many genomic features (Fig. 2-9C) showed a pattern of HIVmGag integration similar to that seen for HIV in Transportin-3- (compare Fig. 2-6) and RanBP2- (data not shown) depleted cells including reduced density of genes, CpG islands, DNase I hypersensitive sites and reduced GC content surrounding integration sites. Thus substitution of HIV gag with MLV gag phenocopied the TNPO3 and RANBP2 knockdowns.
Figure 2-9 – A chimeric derivative of HIV containing MLV gag (HIVmGag) shows reduced integration frequency in gene dense regions. (A) Genetic map of HIV proviruses containing wild type gag (HIVPuro) or a chimera encoding MLV Gag (MA, p12, and CA) in place of HIV MA and CA (HIVmGag). Both viruses have inactivated vpr and env and a puromycin selectable marker in place of nef. (B) Histogram indicating distribution of HIVPuro and HIVmGag integration sites with respect to gene density measured in 1Mb intervals surrounding integration events. Data is plotted as in Fig. 2-4A. Curves are computed using Gaussian kernel density estimates. (C) Genomic heatmap of HIVpuro and HIVmGag datasets. Significant differences are shown by asterisks (*p<0.05; **p<0.01; ***p<0.001).
Knockdowns of RANBP2 or TNPO3 do not cause HIV to favor integration near transcription start sites.

A model to explain the altered integration site patterns of HIV in TNPO3 or RANBP2 knockdowns is that in the absence of these pore proteins the HIV PIC accesses chromatin during nuclear breakdown during mitosis. MLV employs such a mechanism for nuclear entry, so we wondered whether the HIV integration site distributions in the knockdowns might resemble the normal pattern for MLV. Using integration datasets generated by Troy, we asked whether HIV integration in cells depleted of Transportin-3 or RanBp2 shows the most characteristic feature of MLV integration, favored integration near transcription start sites (Fig. 2-10). We found that HIV in the knockdowns disfavors transcription start sites, paralleling HIV integration in unmodified cells. MLV showed strongly favored integration in transcription start sites in the 293T cells studied, and in 293T cells knocked down for TNPO3. We conclude that obstructing the normal HIV pathway of integration by knocking down RANBP2 or TNPO3 does not result in an MLV-like integration targeting pattern. This is consistent with the observation that IN is the dominant determinant of MLV like integration patterns at transcription start sites for chimeric viruses where HIV IN is replaced with MLV IN [49].
Figure 2-10 – HIV and MLV integration patterns at transcription start sites are unaffected by knockdown of TNPO3 or RANBP2. The percent of integration sites within the indicated genomic distances (kb) from the transcription start site (RefSeq genes) is plotted for each dataset. Sample names indicate the VSVG-pseudotyped viral vector used (HIV or MLV) followed by the cell treatment (either control siGL2 or gene-specific siRNA used).

Discussion

Here we report that depletion of Transportin-3 and RanBP2 by RNAi affects the downstream choice of targets for HIV DNA integration, providing evidence for coupling of the nuclear translocation and integration steps. As others have noted, Transportin-3 has little or no effect on infection efficiency of MLV [6,7,9], which is not thought to traverse the nuclear pore, and we report that Transportin-3 did not affect integration targeting by MLV. Replacing HIV gag with MLV gag phenocopied
the effects of the Transportin-3 and RanBP2 knockdowns on HIV integration targeting. These findings support a model in which HIV Gag proteins interact with Transportin-3 and RanBP2 to mediate HIV integration targeting to chromosomal regions rich in genes and associated features.

We found that depletion of several additional factors previously shown to be required for efficient integration also resulted in HIV integration targeting patterns similar to those seen in Transportin-3 and RanBP2 depleted cells. These factors include a component of the anaphase promoting complex (ANAPC2) splicing factors (SNW1 and PRPF38), a WD-repeat protein (WDHD1), and nuclear DNA binding proteins (IK and SNW1). The analysis of some of these was complicated by cell toxicity, and in some cases conflicting results were obtained with different siRNAs, so effects of these factors are less well supported than those of Transportin-3 and RanBP2. It is possible that each of these factors acts in a common pathway with Transportin-3 and RanBP2 to direct integration to regions dense in genes and associated features, though depletion of some of these factors could also alter the synthesis or function of other factors acting more directly.

Our studies support the hypothesis that nuclear import of HIV is linked to integration, and suggest that normal interactions with the nuclear pore help to determine integration target site distributions (Fig. 2-11). We favor a two-step model, in which passage through the pore first places the PIC in regions of high gene density, and then LEDGF/p75 tethers the PIC for integration to provide the final distribution in active transcription units. Several studies suggest that chromosomes
and genes are nonrandomly distributed in the nucleus, though the organization is not fully clarified [51-53]. Although the nuclear periphery is thought to be rich in heterochromatic chromosomal regions that promote gene silencing, studies in yeast and Drosophila suggested that genes can relocate to the nuclear pore upon transcriptional induction [54-59]. Thus passage through the pore may deliver HIV to locally concentrated active gene-dense chromatin. Alternatively, interaction with Transportin-3 and RanBP2 at the pore might engage a nuclear transport system leading to gene-dense chromatin.

Our data is consistent with the idea that correct engagement of the Transportin-3/RanBP2-dependent targeting pathway leads to efficient integration in chromosomal regions rich in genes and associated features. Failure to engage this pathway results in targeting to less gene dense regions. Two possible scenarios can be imagined for nuclear entry and integration targeting in cells depleted for pore factors Transportin-3 and RanBP2.

The first model is that in the absence of Transportin-3 or RanBP2, nuclear access of HIV is restricted to times of nuclear envelope breakdown during cell division. The shift in integration away from gene-dense regions in the TNPO3 and RANBP2 knockdowns may thus reflect changes in chromatin availability during mitosis or shortly afterwards. Consistent with this idea, the HIVmGag virus requires nuclear envelope breakdown during mitosis for infection [50], and it phenocopied HIV integration in the knockdown cells, showing reduced integration frequency in gene dense regions.
An extreme version of this model would hold that HIV integration targeting in TNPO3 and RANBP2 knockdowns might mimic MLV targeting because in both cases the virus accesses chromatin during nuclear breakdown. However, MLV strongly favors integration near transcription start sites, and this is not seen for HIV in knockdown cells (Fig. 2-10).

Similarly, if passage through the nuclear pore delivers the HIV PIC to transcription units and gene dense regions, growth arrest of cells might increase favoring of these features, since all integrants must enter through the pore in arrested cells. Integration site distributions have been investigated in growth arrested IMR90 lung fibroblasts [60] and macrophages [61]. In IMR90 cells, arrest did result in more integration in transcription units and gene dense regions, but in macrophages the favoring is in fact weaker than that observed in many other cell types. Thus it is possible that passage through the nuclear pore results in favored integration in gene dense regions, but additional assumptions are needed to explain the data from macrophages.

The second model (not exclusive of the first) holds that in cells depleted of TNPO3 and RanBP2, HIV integration complexes may pass through the pore but on a different pathway, interacting with different pore proteins. The idea that alternative pathways through the pore exist is supported by findings of Lee and colleagues, who found that the N74D substitution in HIV CA disrupted normal interactions with Transportin-3 and RanBP2 but created dependence on other pore proteins [34]. From our data, it is not possible to determine whether in cells depleted of
Transportin-3 and RanBP2 HIV integration complexes pass through the pore on alternate pathways, or whether nuclear access during mitosis fully explains the data. Thus, at the completion of this study, it was of interest to analyze targeting when integration complexes pass through the pore on alternative pathways, as in the presence of the N74D CA substitution, and such studies are the subject of Chapter 3 of this dissertation.

Acknowledgements

We are grateful to members of the Bushman laboratory for help and suggestions. We would also like to thank Martha Kienzle and members of the Collman lab for help with monocyte isolation and MDM generation; Jianxin You for use of her microscope, advice, and reagents; the University of Pennsylvania CFAR for ELISAs, virus preparations and suggestions; and Paul Bates for reagents.
Figure 2-11 – Model for coupling of nuclear import and integration targeting. Interaction with Transportin-3 and RanBP2 shuttles the PIC through the nuclear pore and toward gene dense regions favored for HIV integration. Interactions with additional factors in the nucleus (ANAPC2, WDH1, IK, PRPF38A, and SNW1) may also play a role in site selection upstream of the known integration cofactor Ledgf/p75, which targets integration to active transcription units. RNA Pol indicates RNA polymerase II, which is known to be required for transcriptional activity, and which promotes integration. Nucleosomes are shown because target DNA is known to be wrapped in nucleosomes during the integration step. PIC, preintegration complex; FG, phenylalanine-glycine repeats of nucleoporins. Adapted from image by Keshet Ronen.
References


CHAPTER 3 – HIV CAPSID INTERACTIONS WITH CYCLOPHILIN-A
AND RAN-BINDING PROTEIN 2 INFLUENCE NUCLEAR IMPORT AND
INTEGRATION PATHWAYS

The contents of this chapter have been published as part of:
Schaller T, Ocieja KE, Rasaiyaah J, Price AJ, Brady TL, Roth SL, Hué S,
Fletcher AJ, Lee K, Kewalramani VN, Noursadeghi M, Jenner RG, James LC,
Bushman FD, Towers GJ. HIV-1 Capsid-Cyclophilin Interactions Determine
Nuclear Import Pathway, Integration Targeting and Replication Efficiency.

Abstract
HIV integrates non-randomly, likely choosing integration sites within the
host chromosomes that best enable the viral genes to be expressed and, ultimately,
progeny virus to be produced. Host factors such as LEDGF/p75 act as cofactors,
promoting integration efficiency and directing integration site selection. We have
reported that depletion of two nuclear pore components, RanBP2 (NUP358) and
TNPO3, alters integration targeting for HIV-1, reducing HIV integration frequency in
gene-dense regions and near gene-associated features. The viral determinant of the
HIV-1 preference for these features maps to HIV gag, leading us to hypothesize that
Gag-dependent engagement of the proper transport and nuclear pore machinery
helps mediate trafficking of HIV complexes to sites of integration. In support of this
hypothesis, isothermal titration calorimetry and other assays demonstrated
functionally that RanBP2 and HIV-1 capsid protein (CA) interact, and the crystal
structure of the cyclophilin-binding loop of HIV-1 CA bound to the nuclear pore
protein RanBP2 shows an interaction that closely resembles that between CA and CypA. To test our hypothesis, we have investigated integration targeting by viruses containing CA mutations that modulate interactions with RanBP2 and/or cyclophilin (CypA). CA mutants interfering with RanBP2 and/or cyclophilin interactions reduced nuclear import of the mutant viruses, and altered the targeting of HIV-1 integration. Mutations on different surfaces of CA either phenocopied or reversed changes seen when RanBP2 or Transportin-3 were depleted, decreasing or increasing the viral preference for gene dense regions and associated features. These findings suggest that direct interactions between HIV capsid and the nuclear pore protein RanBP2 are required for nuclear import of the HIV pre-integration complex and integration in gene-rich regions of the genome. This work may suggest approaches for therapeutic inhibition of these processes.

**Introduction**

The host peptidyl-prolyl *cis-trans* isomerase Cyclophilin A (CypA) binds an exposed Ala-Gly-Pro motif [1] in the amino-terminal domain of HIV-1 CA (amino acids 88-90) and is packaged into mature virions [2,3]. It’s function is a topic of debate, however, the importance of the interaction is supported by the conservation of the exposed loop across primate lentiviruses [4] and the inefficiency of early steps of HIV replication in its absence [5]. CypA has been suggested to act as a chaperone in CA folding [6,7], to destabilize the capsid multimers for efficient uncoating [8,9], and to protect capsid from detection by cellular innate immune
defenses [4,10]. The CA-CypA interaction is inhibited by the immunomodulatory drug cyclosporine.

Interestingly, Ran-binding protein 2 (RanBP2, discussed in Chapter 2), a large nuclear pore protein with multiple functional motifs, contains a cyclophilin (Cyp) domain that catalyzes peptidyl-prolyl cis–trans isomerization of cellular proteins [11,12]. This raised the question of whether RanBP2 might bind HIV CA in a manner similar to CypA. Such an interaction would be consistent with our finding that the preference of HIV for integration in gene dense regions required both RanBP2 and HIV Gag. Moreover, it would suggest another viral step could be functionally coupled in the path to integration, that of capsid disassembly or uncoating.

Work from other groups has also implicated capsid in interactions with the nuclear pore. For example, the HIV-1 CA mutant N57A has a replication defect that is exaggerated in cell-cycle arrested cells suggesting a block at nuclear entry in the absence of nuclear envelope breakdown [13]. In addition, Lee and colleagues isolated a CA mutant (N74D) that bypassed the requirement for Transportin-3 and RanBP2, but acquired a requirement for other nuclear pore factors, suggesting that CA interactions determine pathways available to the PIC for passage through the nuclear pore [14].

Prior to our involvement in this work, Leo James found that the Cyp domain of RanBP2 (RanBP2Cyp) indeed binds to HIV-1 CA with an affinity approaching (within 3 fold) the affinity of CypA for CA. Crystal structures of CA complexed with
CypA and RanBP2Cyp were remarkably similar; however, cyclosporine (Cs), known to inhibit the CA-CypA interaction, did not abrogate binding of RanBP2Cyp to CA. Torsten Schaller, a member of the Towers lab, confirmed an *in vivo* interaction between the RanBP2 Cyp domain and CA: a fusion of the owl monkey restriction factor Trim5 to the Cyp domain of RanBP2 inhibited HIV-1 infection in culture. This interaction was specific to HIV-1 CA, as SIVmac239 CA did not bind RanBP2 Cyp domain and had very low affinity for the CypA Cyp domain. Substitutions at two residues in the conserved CypA binding loop, P90A and G89V, abolished binding to RanBP2Cyp, and decreased (P90A) or abolished (G89V) binding to CypA. Other mutations located at the CA monomer or hexamer interfaces (N57A and N74D), distal to the CypA binding loop, also impaired binding to the RanBP2Cyp. These mutants retained the ability to bind CypA in a manner that was sensitive to cyclosporine. These findings are summarized below, and elaborated in the paper noted at the beginning of this chapter [15].

In this chapter we investigate the role of CA-CypA and CA-RanBP2 interactions in targeting of the HIV-1 pre-integration complex (PIC) to chromatin integration sites in HeLa cells. Specific inhibition of each interaction shows that CypA and nuclear pore proteins have opposing effects on HIV proviral distributions, suggesting that they distinctly influence the path of the HIV PIC to the host chromatin. We discuss possible mechanisms of these findings and implications.
Methods

For detailed descriptions of cell culture, infection and isothermal titration calorimetry methods used by the James and Towers labs, please refer to the publication noted at the beginning of this chapter [15].

Integration site amplification

Parallel preparations of purified genomic DNA were digested overnight one with Msel and the other with NlaIII, ligated at 16°C to PCR adapters, and digested a second time with Sacl. Barcoded nested PCR was then performed using primers and conditions described previously [16-18]. Amplification products between 200-600bp were gel-excised, purified, and sequenced on a Genome Sequencer FLX Titanium Series (Roche 454 Sequencing) at the University of Pennsylvania. Oligonucleotides are listed in Table 3-1. Integration datasets used in this paper are listed in Table 3-2.

Analysis

Integration datasets were analyzed as in Chapter 2 and in previous works [19,20]. For each integration site three matched random control sites (MRCs) were computationally generated that were the same distance from an Msel or NlaIII restriction site. Methods used to determine associations with multiple genomic features are summarized in Appendix 1 of this dissertation. Gene expression
analyses utilized data from HeLa cells measured using the Affymetrix HU133A microarrays as in [21].

Entrez Gene ID Numbers for Genes in this study: RANBP2: 5903, TNPO3: 23534.

<table>
<thead>
<tr>
<th>Table 3-1 - DNA Oligonucleotides used in this study.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Oligo</strong></td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
<tr>
<td>Mse linker+</td>
</tr>
<tr>
<td>Mse linker-</td>
</tr>
<tr>
<td>L1</td>
</tr>
<tr>
<td>Nla linker+</td>
</tr>
<tr>
<td>Nla linker-</td>
</tr>
<tr>
<td>MKL-3</td>
</tr>
<tr>
<td>GW-3</td>
</tr>
<tr>
<td>B-HIV</td>
</tr>
<tr>
<td>shRNA</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>+</td>
</tr>
<tr>
<td>+</td>
</tr>
<tr>
<td>+</td>
</tr>
<tr>
<td>shRNA</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>-</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>control shRNA</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>RANBP2</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
<tr>
<td>TNPO3</td>
</tr>
</tbody>
</table>

**Results**

**Lentiviral CA determines interaction with RanBP2**

The work presented below was collaboration with Torsten Schaller in the Towers lab, and Troy Brady and Shoshannah Roth in the Bushman lab. Torsten generated HeLa cells stably expressing shRNAs to deplete RanBP2 or Transportin-3
and showed that infections by VSV-G pseudotyped GFP-expressing vectors derived from HIV-1 group M (NL4.3) and group O viruses were inhibited in these cells. Replication of intact HIV-1$_{NL4.3-gfp-ires}$ in knockdown HeLa cells expressing CD4 was also impaired. By contrast, SIVmac239 vectors were sensitive to Transportin-3 depletion but not RanBP2-depletion. Given Leo James' finding that SIVmac239 CA does not bind the RanBP2 Cyp domain, this is consistent with the idea that interactions between this domain and CA mediate the HIV requirement for RanBP2. This was further confirmed by CA swap experiments: SIVmac239 bearing HIV CA (SIV-HCA) was sensitive to RanBP2 depletion while group M HIV-1 bearing SIVmac239 CA (HIV-SCA) was insensitive.

**Substitutions in HIV-1 CA Alter Engagement of RanBP2 and Transportin 3**

The hypothesis that a specific CA-RanBP2 interaction mediates the requirement for RanBP2 predicts that HIV bearing CA mutations that prevent interaction with RanBP2Cyp should infect cells inefficiently, and that any infection by these mutants should be insensitive to depletion of RanBP2. Indeed, Torsten found that cyclophilin-binding loop mutants, P90A and G89V, showed reduced infectivity that was not further impaired by RanBP2 depletion. These viruses remained mildly sensitive to Transportin-3 depletion. The N57A and N74D mutants, which are known to retain infection in cycling cells [13,14] despite a loss of interaction with RanBP2Cyp (above), were insensitive to depletion of either RanBP2
or Transportin-3. The data generated by Leo James and Torsten Schaller on the CA mutants is summarized in Figure 3-1 and Table 3-3.

**Fig 3-1 – Crystal structure of HIV-1 CA hexamer.** Hexamer viewed (A) from external surface of capsid and (B) from hexamer interface. CA monomers are colored in alternating shades of green. Residues of interest are highlighted on two adjacent monomers. Mutation of the cyclophilin-binding loop residues Pro90 and Gly89, in red, abolish interactions with Cyclophilin A and RanBP2 cyclophilin domain while alteration of the CA interface residues Asn57 and Asn74, in blue, selectively abolish interaction with the RanBP2. See accompanying Table 3-1. Figure was made using DeepView/Swiss PDB Viewer from PDB entry 3H4E (Pornillos et al. [1]).

**Table 3-3 – Summary of HIV-1 CA mutant interactions with host factors**

<table>
<thead>
<tr>
<th>CA Mutant</th>
<th>Description</th>
<th>CypA Kd (uM)</th>
<th>Nup358 Kd (uM)</th>
<th>Sensitive to RANBP2 KD?</th>
<th>Sensitive to TNPO3 KD?</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td></td>
<td>7</td>
<td>12</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>N57A</td>
<td>monomer interface</td>
<td>7</td>
<td>55</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>N74D</td>
<td>hexamer interface</td>
<td>7</td>
<td>95</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>G89V</td>
<td>CypA binding loop</td>
<td>Nonbinding</td>
<td>Nonbinding</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>P90A</td>
<td>CypA binding loop</td>
<td>46</td>
<td>Nonbinding</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>SCA</td>
<td>SIV CA</td>
<td>Nonbinding</td>
<td>Nonbinding</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>
With Troy Brady and Shoshannah Roth, we identified proviral distributions resulting from infection under several conditions expected to prevent interaction with RanBP2. Given data presented in Chapter 2 showing that depletion of RanBP2 and Transportin-3 led to a decreased HIV integration targeting to features associated with gene density, we expected that mutations abolishing interactions of HIV CA with RanBP2 would affect proviral distributions in a way that would phenocopy the knockdowns.

Integration sites of wild type HIV in knockdown cells, as well as of the mutant viruses and HIV-SCA, are shown with respect to several genomic features in Figure 3-2A. Integration in HeLa cells stably depleted of RanBP2 and Transportin-3 was less biased towards chromatin rich in genes and associated features than in wild type cells, confirming our findings in transient knockdowns of HEK 293T cells (Chapter 2). Similarly, proviral distributions of viruses carrying the N57A and N74D CA mutations were shifted towards regions of lower gene density. However, HIV-SCA and the cyclophilin-binding loop mutants G89V and P90A showed a striking hyper-biased integration targeting pattern, with increased preference for areas rich in genes. Features correlated with high gene density – CpG islands, active transcription units, DNase 1 sensitive sites, and high GC content – were also more highly favored for integration by these viruses.

Hierarchical clustering of the CA mutants based on these data separated the viruses into two groups: N57A and N74D, and the Cyp-binding mutants G89V, P90A and chimeric HIV-SCA (Fig 3-2B). Wild type HIV-1, which has an intermediate
targeting phenotype, was an outlier within this second group. Thus amino acid substitutions in CA can alter integration targeting preference, resulting in either of two phenotypes. Because G89V and P90A influence targeting in the same direction, we infer that disruption of normal CypA interactions results in increased frequency of integration in regions with high densities of transcription units. The N74D and N57A substitutions are less sensitive to depletion of both RanBP2 and Transportin-3, and N74D gains sensitivity to depletion of other nuclear pore proteins [14]. We therefore infer that pathways through the pore that do not involve RanBP2 and Transportin-3 lead to integration in regions with lower densities of genes.

Integration Targeting sensitivities of HIV-1 CA Mutants reflect dependence on RanBP2 and Transportin-3

Depletion of RanBP2 did not significantly affect integration patterns of the mutants except for a slight decreased favoring of GC rich chromatin (Fig. 3-3). This was expected, given that the mutants do not engage RanBP2:CypA and their infection efficiency is not reduced when RanBP2 is depleted. By contrast, proviral distributions of the loop mutant G89A remained sensitive to Transportin-3, mirroring the mutant's sensitivity to TNPO3 knockdown in infection efficiency assays. This observation suggests that Transportin-3 helped target the integration of the G89V mutant, even in the absence of interaction with RanBP2. Proviral distributions of the N74D and N57A mutants were only slightly affected by TNPO3
knockdown, consistent with their relative insensitivity to the knockdown in infection efficiency assays.

**CA-Cyclophilin A interactions influence integration site targeting**

The finding that proviral distributions of viruses bearing CA unable to bind to CypA (P90A, G89V, and HIV-SCA) were shifted to chromatin regions higher in gene density and associated features suggested that CypA-CA interactions might participate in integration site targeting. To test this hypothesis we investigated integration site targeting of wild type HIV-1 in the presence of cyclosporine (Cs), which impairs HIV-1 infection efficiency and selectively inhibits the CA-CypA interaction, without affecting interactions with RanBP2. We found that cyclosporine retargeted HIV-1 integration, phenocopying the hyperbiased targeting of HIV-SCA and the CA loop mutant viruses (Fig. 3-4).

Notably, Torsten found that while Cs treatment and RanBP2 depletion individually reduced wild type HIV-1 infection (as measured by GFP expression from the vector), Cs treatment in the knockdown cells restored infection efficiency to levels in Cs treated wild type cells, suggesting RanBP2 is not required in the absence of CypA and that CypA may act upstream of RanBP2 to determine PIC pathways.
**Figure 3-2** - Effects of CA mutants and knockdowns on integration frequency near multiple chromosomal features. (A) Heatmap showing integration site distributions with respect to multiple chromosomal features. Columns show data for different viruses (HIV-1 bearing SIIMac239 capsid, SIV-SCA). All infections were in wild type HeLa cells except knockdowns indicated. Genomic features analyzed are shown in rows and labeled on the left. Relationships between integration frequency and feature density are summarized using ROC curve areas relative to matched random control distributions. The wild type data-set was used for pairwise statistical comparisons (overlay dashes). P values summarizing the significance of the departure from the WT control are shown with asterisks (*P < 0.05; **P < 0.01; ***P < 0.001). WT virus in TNPO3 and RanBP2 knockdowns as well as N57A and N74D mutants shift to less gene dense regions. SIIMac239 capsid swap, G89V and P90A mutations have the opposite effect. (B) Hierarchical clustering of the integration site sets based on data in (A).
### Figure 3-3 – Effects of knockdowns on integration site distributions of CA mutants.

Heatmaps show integration site distributions with respect to multiple chromosomal features for viruses and shRNA knockdowns indicated as in Figure 3-2. HeLa cells expressing a negative control shRNA (shCtrl) were used for pairwise statistical comparisons for each virus (overlay dashes). P values summarize significance of departure from controls (*P < 0.05; **P < 0.01; ***P < 0.001). Transportin-3 depletion causes broad changes in integration site patterns selectivity for G89V mutant. RanBP2 knockdown causes comparatively little change. Note that for N57A and N74D, integration site distribution changes in knockdowns are modest in magnitude.
**Mutant Integration Patterns are not sensitive to Cyclosporine**

The above integration data is consistent with the idea that CypA and the nuclear pore proteins have opposing effects on integration targeting with CypA acting upstream of RanBP2. Thus we predicted that preventing interactions of the RanBP2 binding-deficient N57A and N74D mutants with CypA by the addition of cyclosporine would lead to hyper-biased proviral distributions of these viruses. The G89V CA does not bind CypA, therefore Cs treatment should not further alter integration targeting of virus bearing this mutation. However, we found that integration sites of all mutant viruses tested were largely unchanged under Cs treatment (Fig. 3-4A). Hierarchical clustering of the datasets based on several genomic features confirmed that, with the exception of wild type HIV-1, the proviral distribution of each virus was most similar to itself regardless of Cs treatment (Fig. 3-4B). This suggests that the N74D and N57A mutants follow a separate pathway into the nucleus that is independent of CypA as well as RanBP2 and Transportin-3.
**Figure 3-4 - Effects of capsid mutations are dominant to that of Cyclosporine treatment.** (A) Heatmap shows integration site distributions with respect to multiple chromosomal features for viruses in the presence or absence of cyclosporine as indicated. Heatmap analysis is assembled as described in Figure 3-2. Wildtype HIV-1 integration sites in untreated HeLa cells were used for pairwise statistical comparisons (overlay dashes). P values summarizing the significance of the departure from the WT control are shown with asterisks (*P < 0.05; **P < 0.01; ***P < 0.001). Cyclosporine treatment did not significantly affect integration targeting for any of the three mutants shown. (B) Hierarchical clustering of the integration site sets based on the data in (A).
Discussion

Leo James and Torsten Schaller demonstrated for the first time a direct interaction between HIV-1 CA and the nuclear pore, via the cyclophilin domain of RanBP2. In this work, we have used deep sequencing to investigate perturbations of viral CA interactions with host factors and their effects on integrated proviral distributions. The data presented suggest that that HIV-1 targets preferred regions of chromatin for integration by means of a pathway that includes both CypA and RanBP2.

Mutations in CA that caused the virus to be less dependent on RanBP2 and Transportin-3 (N57A and N74D) also caused the virus to integrate in regions of chromatin less dense in genes and associated features. This phenotype mimics integration patterns of wild type virus in RanBP2 and Transportin-3 depleted cells, confirming the role of these factors discussed in Chapter 2. The pathway taken by these mutant viruses to the chromatin also appears to be CypA independent. By contrast, mutating the viral CA such that the virus could not interact with CypA or RanBP2 (P90A and G89V) led to infection and integration by a pathway that was independent of RanBP2 but still sensitive to Transportin-3. Strikingly, blocking interaction of CA with CypA, either by mutation or treatment of wildtype-HIV-1 infected cells with Cs, led to increased frequency of integration in chromatin areas with higher densities of transcription units and associated features, an effect opposite to that of selectively preventing CA-RanBP2 interactions. This data suggests that CypA moderates the natural HIV-1 targeting bias, placing the PIC on a
pathway with decreased preferences for gene density and associated features, some of which are restored by nuclear pore proteins. Alternatively CypA may reduce the availability of RanBP2 to the HIV PIC.

However the role of CypA remains obscure. Why do some viruses that cannot interact with CypA - the G89V and P90A viruses, and wild type virus in Cs-treated cells – take one integration pathway targeting gene dense regions, while the N57A and N74D viruses target chromatin less enriched for genes, even in the presence of Cs? While the mechanism has not been determined, we infer that the N57A and N74D mutations cause changes in CA that allow interactions with additional host factors to enable entry to the nucleus and integration (Fig 3-5). In fact, it has been shown that the N74D virus gains a new dependence on the nuclear pore protein Nup155 [14], and the observed integration patterns support proposals that there are multiple pathways into the nucleus. In some cases, as for the N74D CA mutant, alternate pathways might be equally efficient for infection of cells in culture [14].

Interestingly, SIVmac239 did not appear to use RanBP2 during infection, although it remained dependent on Transportin-3. This implies that SIVmac239 uses a slightly different set of factors to transport its PIC into the nucleus. While observed patterns of SIV and HIV integration are largely similar [22,23], subtle differences may exist. It will be of interest to investigate the pathways used by a variety of lentiviruses to traverse the nuclear pore and achieve preferred proviral distributions.
We hypothesize that CA interactions help the HIV-1 PIC to select co-factors for nuclear import and integration site targeting. HIV-1 may have evolved to communicate with several host factors to achieve its preferred final distribution of integration sites. While we have not observed differences in HIV expression upon retargeting integration [24], it may be that integration targeting subtly calibrates gene expression with biological significance in vivo [25]. The interaction reported
here between HIV-1 CA and RanBP2 could occur in the context of a PIC containing CA molecules as suggested by nuclear targeting studies [13,26], or with an intact or partially intact capsid, consistent with electron microscopy images published in one report showing capsid cores docked at the pore [27]. Engagement of an intact capsid by several RanBP2 molecules arranged at the pore might be thermodynamically distinct from interactions with soluble CypA in the cytoplasm and could lead to destabilization of the capsid, releasing the PIC to factors such as Transportin-3. A link between uncoating and nuclear import has been previously suggested [28]. This mechanism might enable protection of the PIC from innate immune factors in the cytoplasm and efficient and appropriate integration targeting in the nucleus. It will be interesting to investigate the behavior of the CA mutants within the hexameric unit (or even polymerized as in the fullerene capsid). Perhaps the N57A and N74D mutations, located at the monomer and hexamer interfaces of the CA protein respectively, alter as-yet uninvestigated interactions of the intact capsid. In addition, the role of cis-trans isomerization by either CypA or RanBP2 at CA residue P90 has yet to be studied. Finally, as evidenced by cyclosporine, it is possible to inhibit cyclophilin domains such as that of RanBP2 from binding their substrates, and thus RanBP2 is a potential target for therapeutics.

**Acknowledgements**

We would like to thank members of the Bushman, Towers and James lab for productive collaboration, helpful suggestions, and stimulating discussion.
References

CHAPTER 4 – INTEGRATION SITE SELECTION BY A HIV89.6, A LOW PASSAGE ISOLATE, IN PRIMARY HUMAN T CELLS

Abstract

HIV-1 integrates into the host genome as a requisite step of its replication. The genomic loci chosen for integration are non-randomly distributed with respect to many genomic features, reflecting interactions with several host co-factors that help target the viral pre-integration complex to the chromatin. It is thought that HIV-1 has evolved to interact with host factors that direct it towards chromatin that is favorable for expression of the viral genome. Here we present a study of HIV-1 integration sites in primary CD4+ T cells, sequenced in unprecedented depth. Our analysis confirms known targeting preferences of HIV-1 for active transcription units and chromatin rich in associated features and epigenetic markers. However, statistical analysis of specific phosphodiester bonds of the human genome that host repeated insertions and of unexpectedly favored regions suggests that additional characteristics of the targeted sites exist, attracting the preintegration complex to insert the HIV-1 genome in those sites.

Introduction

Retroviruses integrate a copy of their genome into the genome of the infected host cell. This essential step of the viral replication cycle occurs in a non-random
fashion, with viruses likely choosing integration sites best suited for expression of their genome [1-3]. This can also affect the host cell, because integration in genes or regulatory elements can cause insertional mutation. Insertional activation by a gammaretroviral vector has led to transformation in several patients treated by gene therapy [4], and in another trial, insertion of a lentiviral vector carrying the β-globin gene disrupted regulation of HMGA2 by micro-RNAs and was associated with clonal expansion [5].

Various retroviruses show different proviral distributions with respect to genomic features and epigenetic marks, reflecting different mechanisms used to target favorable integration sites. For example, while MLV prefers to integrate in promoters and near gene 5’ ends, HIV-1 integrates along the length of active transcription units, also showing a preference for nucleosome-bound DNA and genomic regions rich in genes, GC-content, CpG islands, and DNase hypersensitive sites [3,6-13]. The HIV-1 distribution is influenced by the human DNA-binding protein, PSIP1/LEDGF/P75 (LEDGF), which binds HIV-1 IN and tethers the pre-integration complex (PIC) to the chromosome for integration [14-16]. It’s role in directing integration site selection is supported by studies in which LEDGF-regulated genes were favored for HIV-1 integration, and in which depletion or retargeting of LEDGF changed HIV-1 integration site distributions [11,17]. LEDGF does not bind MLV IN, nor does it affect MLV integration site selection [18]. Highly active inhibitors targeting the LEDGF binding site on HIV-1 IN are now available, and their effects on integration site selection during HIV-1 infection are under-
studied. CypA and components of the nuclear pore not engaged by MLV may also be involved in a pathway directing HIV-1 integration to favorable gene-dense chromatin as discussed in Chapters 2 and 3 [19,20].

Our knowledge of integration preferences of HIV-1 is largely derived from proviral patterns of HIV-1-based vectors in cell lines. Several studies suggest that these patterns likely reflect integration targeting biases during human HIV-1 infections: although subtle differences in integration site targeting is seen between cell types (for example in macrophages or between resting and active T cells) the proviral distributions of different HIV-1 vectors in all cell types tested remain essentially similar [6,8,10,12,13,21-23]. However, no large-scale study of integration sites from a low passage HIV-1 isolate in human T cells has been performed.

Increasingly large datasets of integration sites afford us finer resolution views of genomic regions preferred for integration and enable us to ask whether the determinants listed above (active transcription, CpG islands, etc.) as well as the known weak sequence specificity of IN fully describe the biases of HIV-1 integration targeting at the chromosomal and local levels. We do not yet fully understand all the determinants of HIV-1 integration site selection – for example, the virus retains some integration preferences in the absence of LEDGF or pore factors. Characteristics distinguishing the most preferred sites for integration could provide clues to unidentified mechanisms.
Here we present what is to date the largest dataset of HIV integration sites – close to 150,000 – generated by a low passage viral strain in primary human CD4+ T cells. The infectious clone shows similar integration site selection preferences to HIV-1 based vectors in cell lines, targeting active transcription units, and associated epigenetic markers. Broad regions rich in genes and correlated features are also targeted, while at the local level, the natural sequence specificity of IN is observed. A major goal of this study was to evaluate our current understanding of integration targeting. Replicate infections investigated in this study – previously impractical to sequence – allowed us to identify specific phosphodiester bonds (“hotspots”) in the human genome that are repeatedly targeted by HIV. Analysis of these genomic loci suggests that preferred genomic features, epigenetic marks, and local sequences combine to make these sites optimal. However, ongoing analysis by our collaborator, Charles Berry, suggests that known biases are unable to fully account for these sites, and current statistical models predicting integration sites based on genomic annotations fail to account for observed variation in viral integration distributions. Thus novel preferences and mechanisms must exist, motivating further research.

**Methods**

**Cell culture and viral infections**

HIV-1 stocks were generated by the University of Pennsylvania Center for Aids Research. 293T cells were transfected with plasmid containing the viral clone,
and harvested virus was passaged in SupT1 cells once. Virus was titered by p24 antigen content. Primary CD4+ T cells were isolated by the University of Pennsylvania Center for AIDS research Immunology Core from apheresis product from a healthy male donor using the RosetteSep Human CD4+ T Cell Enrichment Cocktail (StemCell Technologies).

T Cells were stimulated for 3 days at 0.5 x10^6 cells per milliliter in R10 media (RPMI 1640 with GlutaMAX (Invitrogen) supplemented with 10% FBS (Sigma-Aldrich) with 100 units U/mL recombinant IL2 (Novartis) + 5ug/mL PHAL (Sigma-Aldrich). Cells were infected in triplicate. For each infection, 6.6x10^6 cells were mixed with 1.32 µg HIV-1^89.6 in a total volume of 2.25 mL. This was split over three wells of a 6 well plate for spinoculation at 1200g for 2hr at 37°C. Cells were incubated an additional 2hr at 37°C. Cells were then pooled into flasks and volume was increased to a total of 12mL. Spreading infection was allowed to proceed 48hr at 37°C, after which cells were harvested. 1x10^6 cells were harvested for flow cytometry, and 6x10^6 cells were pelleted following two washes in PBS for nucleic acid extraction (below).

**Flow cytometry**

To assess percent infected cells, 1x10^6 cells were stained per infection for flow cytometry. All incubations were at room temperature. Cells were first washed in PBS and then twice in FACS wash buffer (PBS, 2.5% FBS, 2mM EDTA). Cells were fixed and permeabilized with CytoFix/CytoPerm (BD) for 20 minutes and washed
with Perm-Wash Buffer (BD) before staining with anti-HIV-Gag-PE (Beckman Coulter) for 60 min. Finally cells were washed in FACS wash buffer and resuspended in 3% PFA. Samples were run on a LSRII (BD) and analyzed with FlowJo 8.8.6 (Treestar). Cells were gated as follows: lymphocytes (SSC-A by FSC-A), then singlets (FSC-A by FSC-H), then by Gag expression (FSC-A by Gag).

Integration site amplification

Genomic DNA was isolated from 6x10^6 T cells per infection using the AllPrep DNA/RNA Mini Kit (Qiagen) with Qiashredder columns (Qiagen) for homogenization according to the manufacturer’s instructions. DNA was eluted in 140 µL elution buffer. Purified genomic DNA was digested overnight with NlaIII and ligated at 16°C to a linker, which served as an adapter for PCR. Each sample was ligated to a different linker containing a unique 40bp sequence within the site bound by PCR primers. Different PCR primers were similarly uniquely matched to each sample. This design was intended to reduce the possibility of PCR contamination between the three samples (Fig 4-1). After adapter ligation, samples were digested a second time with BglII. Barcoded nested PCR was then performed in 12 replicates per sample with primers complementary to the 3’ HIV-1_{1896} LTR and the unique ligated linker much as described [6,22,24]. The following conditions were used for the first round of PCR: 60s at 94°C; 7 cycles of 2s at 94°C and 60s at 70°C; 37 cycles of 2s at 94°C and 60s at 67°C; and finally 4 minutes at 72°C. Second round PCR adding adapters and barcodes for 454 sequencing was performed as
follows: 60s at 94°C; 5 cycles of 2s at 94°C and 60s at 70°C; 20 cycles of 2s at 94°C and 60s at 67°C; and 4 minutes at 72°C. Amplification products between 200-600bp were gel-excised, purified, and sequenced on a Genome Sequencer FLX Titanium Series (Roche 454 Sequencing) at the University of Pennsylvania. Oligonucleotides are listed in Table 4-1.

![Diagram](image.png)

**Figure 4-1 - Minimizing PCR cross-contamination in amplifying integration sites.** Genomic DNA (blue) containing integrated proviral DNA (red) was isolated from infected cells and digested with the restriction enzyme NlaIII. Sticky ends were ligated to linkers which serve as adapters for PCR. A different linker containing a unique 40 basepair sequence (orange, purple, pink) was used for each sample. Nested PCR was performed with primers recognizing the HIV-1 LTR and the unique linker adapters, such that each PCR reaction could only amplify integration sites from the matched linker-ligated sample. The second round of PCR added barcodes (different for each sample, orange, purple, pink) and 454 adapters (turquoise and green) allowing sequencing from the 3’ LTR into the adjacent host human genomic DNA.
Table 4-1 - DNA oligonucleotides used in this study.

<table>
<thead>
<tr>
<th>Oligo</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nla linker+</td>
<td>zzzzzzzzzzzzzzzzzzzzeeeennnnnnnnnnnnnnnnCTCGCTTAAGGGACCATG</td>
<td>Linker oligo plus strand, z's, e's and n's: 40bp sequence unique to each linker</td>
</tr>
<tr>
<td>Nla linker [Phosp]GTCCCTTAAGCGGAG-[AmC7-Q]</td>
<td></td>
<td>Linker oligo minus strand, 5'phosphate and 3' amide C7 modification</td>
</tr>
<tr>
<td>PCR1-L</td>
<td>zzzzzzzzzzzzzzzzzzzzeee</td>
<td>Linker primer for first round PCR (z's and e's from matched linker)</td>
</tr>
<tr>
<td>PCR2-A-L</td>
<td>GCCTCCCTCGCGCATCAGeeenennnnnnnnnnnnnnnnn</td>
<td>Linker primer for second round PCR (e's and n's from matched linker)</td>
</tr>
<tr>
<td>MKL-3</td>
<td>CTAAAGCTCAATAAAGCTTGCTTGAG</td>
<td>HIV LTR primer for first round PCR</td>
</tr>
<tr>
<td>Barcoded B-HIV896</td>
<td>GCCTTGCCAGCCCGCTCAGXXXXXXXXXAGACCATCCTAGTTTAAGTTAGAAAATCTC</td>
<td>HIV LTR primer for second round PCR (Xs: barcode for 454 sequencing)</td>
</tr>
</tbody>
</table>

*Second round PCR primers add A and B adapters for 454 sequencing

Analysis

Integration site datasets used in this study are listed in Table 4-2. Each dataset was individually collapsed to the set of it’s unique integration sites and filtered: genomic sequences that began within three base pairs of the LTR end and showed unique best alignments to the human genome by BLAT (hg18, version 36.1, >98% match score) were considered true integration sites. The NlaIII restriction enzyme leaves 5'-CATG-3' overhangs. Because the proviral LTR terminates in a 5'-CA-3' sequence, any integration event adjacent to a genomic 5'-TG-3' can be cleaved in the NlaIII digestion used to fragment the genome in order to amplify integration sites. During the ligation step, cleaved proviral ends from such integration sites can
re-ligate to digested genomic DNA, artificially randomizing these integration sites. Additionally, abortive circular and linear forms of the unintegrated virus can form CATG sequence at their LTRs. To eliminate these artifacts, datasets were filtered to remove integration sites found within a host genomic CATG sequence or within 2bp of such a sequence.

Genomic loci (specific phosphodiester bonds on the same DNA strand) called as integration sites for more than one infection (in two of three or all three replicate infections) were pooled into an additional dataset of “hot sites” for analysis. The repeated sites were not dereplicated within these sets so as to weigh the sites according to the number of times each was observed, a measure of their attractiveness for integration. A reciprocal set of pooled unique integration sites each seen in only one infection was also created for analysis.

Integration sites inferred to have integrated at the same genomic location but in the opposite orientation were identified as sequences mapping to the same genomic location but on the opposite genomic DNA strand, taking into account the fact that the two viral ends are offset by 5bp during the integration reaction resulting in the replication of those bases on either side of the virus (such that an integration site on the target +strand matches one 5bp downstream on the –strand with respect to the host chromosome).

Comparisons to genomic features were carried out as described previously [21,25] using a combination of conditional logit regression and Bayesian model averaging. For each integration site seven matched random control sites (MRCs)
were computationally generated that were the same distance from an NlaIII restriction site as the corresponding experimental site but otherwise randomly distributed in the human genome. These sites were filtered for CATG proximal sites as above. These sites were used to control for biases in isolating integration sites due to uneven distributions of the NlaIII restriction site. Details of statistical methods are available in [21]. Methods used to determine associations with multiple genomic features are summarized in Appendix 1 of this dissertation. Gene expression analyses utilized data from active T cells [26] with expression measured using the Affymetrix HU133 plus 2.0 gene chip array.

Table 4-2 - Integration site datasets used in this study

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Virus</th>
<th>Cell Type</th>
<th>Restriction Enzyme*</th>
<th>Number of Sites</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infection I Integration Sites</td>
<td>HIV-89.6</td>
<td>CD4+ T</td>
<td>NlaIII</td>
<td>56500</td>
<td>This work</td>
</tr>
<tr>
<td>Infection II Integration Sites</td>
<td>HIV-89.6</td>
<td>CD4+ T</td>
<td>NlaIII</td>
<td>48170</td>
<td>This work</td>
</tr>
<tr>
<td>Infection III Integration Sites</td>
<td>HIV-89.6</td>
<td>CD4+ T</td>
<td>NlaIII</td>
<td>42630</td>
<td>This work</td>
</tr>
<tr>
<td>Integration sites repeated in more than one infection</td>
<td>HIV-89.6</td>
<td>CD4+ T</td>
<td>NlaIII</td>
<td>1956 (960 unique)</td>
<td>This work</td>
</tr>
<tr>
<td>Integration sites observed in only one infection, pooled</td>
<td>HIV-89.6</td>
<td>CD4+ T</td>
<td>NlaIII</td>
<td>145344</td>
<td>This work</td>
</tr>
<tr>
<td>HIV Vector/Cell line sites</td>
<td>VSVG-GFP-HIV vector</td>
<td>Jurkat-T</td>
<td>Msel</td>
<td>21055</td>
<td>Wang et al. 2007</td>
</tr>
<tr>
<td>HIV Vector/Cell line sites</td>
<td>VSVG-GFP-HIV vector</td>
<td>Jurkat-T</td>
<td>AvrII, Spel, Nhel</td>
<td>23181</td>
<td>Wang et al. 2007</td>
</tr>
<tr>
<td>MLV Vector/Cell line sites</td>
<td>VSVG-GFP-MLV vector</td>
<td>HEK-293T</td>
<td>Msel</td>
<td>616</td>
<td>Ocwieja et al., 2011</td>
</tr>
</tbody>
</table>

*Enzyme used to digest genomic DNA in order to recover integration sites
Results

Distribution of HIV-1\textsubscript{89.6} integration sites in the human genome

To investigate large replicate datasets of HIV-1 proviral distributions, we amplified integration sites from three independent infections of stimulated primary CD4+ T cells from a single human donor, infected \textit{in vitro} with HIV-1\textsubscript{89.6}. The percent infected cells varied, with 37.5\%, 26\% and 21\% infected in cultures I, II and III respectively (not shown) and this was reflected in the number of integration sites recovered (Table 4-2). As detailed in the methods and illustrated in Figure 4-1, a labeled linker approach was used to minimize the possibility of small amplicons from one sample contaminating another during amplification of integration sites.

We first plotted the integration sites from each infection along the human chromosomes (Fig. 4-2). As has been previously observed \cite{6}, HIV-1\textsubscript{89.6} targeted the chromosomes unequally for integration, even after normalizing for the different chromosome lengths, with more integration sites on those chromosomes more dense in genes (for example Chromosomes 17 and 19) (Fig. 4-2A). Figure 4-2B shows HIV-1\textsubscript{89.6} integration sites, plotted using the UCSC genome browser (hg18, http://genome.ucsc.edu) along selected 100 kilobase regions of human Chromosomes 19 and 14, which were especially attractive and relatively disfavored for HIV-1 integration respectively. For comparison, we also show sites isolated in an earlier large study of HIV-1 vector integration in Jurkat T cells \cite{6}. One observes that distributions of sites are largely similar among multiple infections and across the studies. While we cannot rule out that the use of different restriction enzymes

107
Figure 4-2 - Integration sites, hot sites, and hot regions. (A) Percent of integration sites in each dataset normalized to percent of matched random control sites (MRCs) on each chromosome (to control for length). Pooled unique sites: integration sites appearing in only one infection; pooled hot sites: insertion sites appearing in more than one infection. (B) Integration sites from three infections and from two different preparations in a previous study (Wang et al. [6]) were plotted on the UCSC Genome Browser (hg18). Shown are two representative 100 kilobase regions on Chromosomes 14 and 19. Hot sites, as in (A) in green. Turquoise bars indicate regions observed to have 20 or 100 integration sites where fewer were predicted by statistical models ("Hot Regions"). UCSC genes and Genbank mRNAs are also shown.
in the two studies led to disparities in our ability to recover sites in some locations, we note subtle differences between the HIV-1\textsubscript{89,6} and HIV-1 vector proviral distributions. In addition, the large number of integration sites in our set adds resolution to the picture – one can observe much local variation that is clarified by the more densely sampled proviral distributions.

**Genomic and Epigenetic Features Associated with HIV-1 Integration**

As an initial survey of the integration targeting preferences of HIV-1\textsubscript{89,6}, we analyzed correlations between integration sites from each of our three replicate infections with several genomic features. In the heat map shown in Figure 4-3A, each correlation is plotted as a tile, colored according to the direction and strength of the correlation (for guidelines to interpreting heat map displays, see Appendix 1 and [21]). HIV-1\textsubscript{89,6} integration showed predictable targeting to regions high in gene density and associated features such as CpG islands. Genomic DNA surrounding integration sites was richer in G/C content than at matched random sites, except in the immediate vicinity of the sites. The local relative increase in AT content is thought to reflect HIV-1’s preference for integration on nucleosomes and the role of LEDGF, which has an A/T-hook DNA-binding motif [6]. As has been previously reported, HIV-1\textsubscript{89,6} showed a distinct preference for integration within genes (Fig. 4-3B), with increased favoring of more actively transcribed genes (Fig. 4-3C). The preferences of HIV-1\textsubscript{89,6} in CD4 T cells were reproducible across infection replicates and quite similar to those observed in a previously published large
integration dataset in which the Jurkat-T cell line was transduced with a VSV-G pseudotyped HIV-1 derived vector (Figure 4-3A) [6]. Integration patterns of HIV-1\textsuperscript{89.6} differed from previously observed biases of MLV (Fig. 4-3A) [13,19].

We used published ChIP-seq data on the location of 39 epigenetic modifications, RNA Polymerase II (PolII), and the transcriptional repressor CTCF CD4\textsuperscript{+} T cells [27,28] to similarly investigate the association of HIV-1\textsuperscript{89.6} integration sites with chromatin markers (Fig. 4-4). As previously reported [17], HIV-1\textsuperscript{89.6} integration was favored in areas high in several epigenetic modifications associated with active transcription, including H2BK5me1, H4K16ac, H4K20me1 and additional histone acetylations found in highly expressed genes in CD4\textsuperscript{+} T cells [28]. Integration was strongly negatively correlated with H3K9 and H3K27 di- and tri-methylations, modifications associated with transcriptional repression [29].

**Phosphodiester bonds repeatedly targeted for HIV-1 integration**

We identified 960 specific phosphodiester bonds of the human genome that were repeatedly targeted in independent integration events ("hot sites"). Some of these are depicted in Figure 4-2. Of these sites, 36 were observed in all three infections. Such hot sites are difficult to distinguish within a single culture because infected cells may divide, passing proviruses to progeny cells, and because integration sites must be PCR amplified for 454 sequencing.

Three lines of evidence suggest that the hot-sites reflect true independent integration events at the same genomic location rather than artifacts of PCR
Figure 4-3 - HIV-189.6 Integration sites compared to distributions of chromosomal features. (A)
Heatmap showing positive and negative correlations with several chromosomal features. Integration
sites from HIV-189.6 infections are plotted and sites of an HIV-1-derived vector in Jurkat T cells (Wang et
al. 2007) and of an MLV-derived vector in HEK-293T cells (Ocwieja et al. 2011) are included for compari-
son (columns). Genomic features analyzed are shown in rows. Relationships between integration
frequency and feature density are summarized using ROC curve areas, relative to matched random
control (MRC) distributions. The Infection II set was used for pairwise statistical comparisons (overlay
dashes). P values summarize significance of departure from Infection II (*P < 0.05; **P < 0.01; ***P <
0.001). Base pair values in row labels indicate the size of the genomic interval used for analysis. (B)
Integration site distributions relative to RefSeq genes and (C) relative gene expression intensity. The
measured value for the insertions in (B) was divided by that of the MRC sites to emphasize the departure
of the experimental data from random which was significant - p value is based on regression analysis.
Figure 4-4 - Integration intensity relative to density of epigenetic markers. Integration site datasets are shown in columns with sites from a previous study of HIV-1 derived vector integration in Jurkat T cells (Wang et al. 2007) for comparison. Histone methylations and acetylations and transcription factors (rows) are grouped into clusters which are observed to colocalize and associate with functional genomic elements and expression in CD4+ T cells (Wang et al 2008). Color tile indicates ROC area based on comparisons between integration sites and matched random controls. Epigenetic marks are measured in a 10kb window around integration sites.
contamination between samples. First, if the repeated sites were derived from mixing of a small amount of amplified material between samples, we’d expect the repeated sites to be a representative subset of the total amplified integration sites, showing similar distributions with respect to genomic annotations. As described below, this was not the case – the hot sites showed statistically different patterns. Second, if the hot sites are truly more attractive for integration, then the odds of finding the same sites in another dataset prepared independently should be greater for the sites found in 2 or more replicate infections than for the sites recovered from only one infection (“unique sites”). We analyzed HIV-1 integration sites reported by Wang et al. [6] – amplified several years prior to this work using different restriction enzymes, PCR adapters and primers – and found the odds of finding the same site in the Wang et al. data was greater for insertions observed in two infections (1.1x10^{-2}) than those observed in only one infection (1.8x10^{-3}), and greater still for integration sites seen in all 3 infections (5.9x10^{-2}). In contrast, the odds of finding a matched random control site from the current study in the Wang et al. insertions were much lower (3.0*10^{-6}). Finally, given what we know about the integration reaction, a specific genomic location should be equally attractive for insertion of the viral cDNA in either orientation with respect to the genome [30,31]. Restriction enzyme sites are not equidistant from either side of insertions, so we could not recover sites in either orientation with equivalent ease; however, the odds of observing the same site in the opposite orientation in our dataset were greater for the proposed hot sites (4.3x10^{-2}) than for the unique sites (7.9x10^{-3}), suggesting that the repeatedly
observed sites are truly more attractive for integration. Our collaborator, Charles Berry, performed the above analysis.

Genomic sites repeatedly targeted for integration are found in regions especially rich in preferred features

Human phosphodiester bonds repeatedly hosting insertions might fall in “perfect storm” locations enriched for all of the features favored by HIV-1. They might alternatively be enriched in specific favored features, or they may be targeted due to novel characteristics. To investigate these possibilities, we compared the biases of the pooled set of repeated “hot sites” to those of sites observed in only one of our three infections. We looked first at the distribution of the hot sites with respect to the human chromosomes. As shown in Figure 4-2A, compared to the uniquely targeted sites, the hot sites are much more likely to be found on the preferred chromosomes and less likely to be found on disfavored chromosomes.

We again investigated both sets of pooled integration sites with respect to several mapped genomic features and epigenetic marks (Fig. 4-5). The positive and negative correlations of HIV-1_{89.6} with respect to genes, gene density, expressed genes, CpG islands, DNase sites, and GC content were all significantly strengthened at the hot sites when compared to the unique sites in our datasets (Fig. 4-5A). Large magnitude differences were seen for CpG density and gene density when measured in 1Mb windows around the integration sites, and G/C content was significantly higher in broad regions surrounding the repeated sites. Interestingly, the local
regions immediately surrounding the hot sites were not significantly different from
the unique sites in terms of their G/C richness, but trended to increased G/C
content. It has been suggested that the relative favoring of A/T DNA locally at
integration sites reflects a preference for integration on nucleosomes and near
binding sites for the A/T hooks of LEDGF [6], but this preference did not appear
dominant.

Positive and negative correlations with most epigenetic features were also
maintained, and were stronger for the hot sites than for the unique sites (Fig. 4-5B).
However, two marks of heterochromatin, H4K20me3 and H3K79me3 [27,32,33],
which are normally disfavored for HIV-1 integration, were selectively less
disfavored at the hot sites and these are discussed further below.

To investigate more closely the preference for integration in genes, we
plotted insertion sites with respect to the transcription start site of the nearest gene
and found again that the hot sites had a distribution that was similar to but more
exaggerated than the unique integration sites (Fig. 4-5C). Gene promoters (favored
by MLV) are disfavored by HIV-189,6 [13] and phosphodiester bonds targeted by
insertions in more than one infection were even less likely to occur in promoters.
Genomic loci short distances from the promoters –especially downstream, in genes
– are preferred by HIV-1 and repeatedly targeted sites are more likely to be in these
regions. The percent of hot sites was lower than the percent of unique insertions
falling in locations unlikely to be in genes, very far (>25kb) from transcription start
sites (not shown).
Figure 4-5 - Distribution of hot sites relative to density of chromosomal and epigenetic markers. Integration distributions plotted in genomic (A) and epigenetic (B) heat maps as in Figures 4-3 and 4-4. MRC, matched random control. Sites observed in only one infection are used for pairwise statistical comparisons and *P values summarize the significance of the departure (*P < 0.05; **P < 0.01; ***P < 0.001). (C) Percent of integration sites within indicated genomic distances (kb) from the transcription start site (RefSeq genes).
Primary sequence at genomic locations repeatedly targeted for integration

The data presented above is consistent with the idea that a globally favorable chromosomal environment promotes integration at the hot sites, but it does not explain the virus's choice of specific phosphodiester bonds. Therefore we looked for altered biases at the local sequence level. Here we found that the 20 base pair sequences surrounding hot sites adhere better to the sequence preferred by integrase with greater information content measured for the hot sites than unique sites at every base (Fig. 4-6). The local sequence displayed the previously described weakly conserved inverted repeat between bases -3 and 7 with respect to the integration site [6,7,21,34-36].

Statistical modeling reveals presence of unknown motivators of targeting bias

Our collaborator, Charles Berry, sought to determine whether the correlations observed previously and here fully describe the integration distribution of HIV-1, accounting for preferred regions and preferred phosphodiester bonds found in the data. His work, still in progress, suggests that we do not yet completely understand the relative importance of the known preferences, or that there are unknown determinants of integration site selection that must be added to his statistical models.

Dr. Berry uses quantitative models of integration preferences to predict integration intensity across the genome. If we were to pick a site at random from our datasets, not knowing whether it was a real insertion or a computationally
Figure 4-6 - Favored local sequence at HIV-1 integration sites. The height of the letter reflects the degree of conservation. Y axis show's bits of information (perfect conservation is 2 bits). Integration occurs between bases 0 and -1 or between bases 4 and 5 on the complementary strand. (A) Unique sites and (B) hot sites are found in similar weakly conserved local sequences with slightly different informatoin content. Plots were created using the WebLOGO program (www.Weblogo.berkeley.edu).

generated matched random control site (MRC), and we were told that the site was in a particularly gene dense region, we might predict it was an insertion. Multiple conditional logit regressions combine into a single predictive rule the complex effects of several such measured features that vary predictably with respect to the probability that a site is an insertion rather than an MRC. The model is corrected for confounding effects of correlated types of genomic annotations. We can then test the model to see how well the predictive rule fits the data. Additional variables
(such as whether a site was targeted in two or more independent infections) can be added, and we can ask how much these variables improve our ability to predict insertions in the data. If the model is perfect, the effects of new variables should be encompassed and predicted by components of the original rule, and the new variables should therefore have coefficients of zero. If the data is best fit by a model in which the new variables have non-zero coefficients, they tell us something the original rule didn’t. This approach is described in detail in a previous work [21].

From the data above, it appears that the hot sites are favored because they fall in locations that are rich in favored features and are poor in disfavored features. However, analysis by Dr. Berry suggests that this is not the entire story – the presence of an insertion at the same location in another infection yields an almost 20 fold increase in the odds that a site is an insertion rather than an MRC after applying the predictive rule based on the genomic features presented above. As follows, this “other infection” variable has a non-zero coefficient when fitting the data to a conditional logit model, suggesting that something unknown is contributing to the preference of these phosphodiester bonds.

Looking at the entire dataset, Dr. Berry identified several regions in which there were 20 or 100 insertions, but where the model predicted there would be fewer. One of these “hot regions” is shown on Chromosome 19 in Figure 4-2 above. When the variables of whether or not a site was in a 20- or 100-insertion hot region were added to a conditional logit model, they had non-zero coefficients. The Z score indicates how much a variable adds to the model when added last, thus serving as a
measure of importance. When comparing the variables tested (the hot regions and repeated sites in either orientation), the 100-insertion hot region variable was most influential. Work is underway to determine what distinguishes the hot sites and hot regions, making them more attractive than we can predict with the genomic features considered.

**Discussion**

Here we present close to 150,000 HIV-1 integration sites, recovered from three replicate *in vitro* infections of primary human CD4+ T cells. Improved 454 pyrosequencing technology made it practical to generate this large dataset, which provides a finer resolution picture of integration site distributions, maximizing our ability to discriminate between favored and disfavored genomic and epigenetic features. Additionally, improved amplification methods reducing the possibility of cross contamination between PCR reactions enabled us to conduct new analyses on genomic loci targeted in replicate infections.

Never before had a large study of integration sites of an intact virus in CD4+ T cells been performed. As we expected, the distribution of the low passage isolate, HIV-189.6, closely resembled that reported for HIV-1-derived vectors in cell lines with a preference for integration in transcription units and in chromosomal regions rich in active genes and correlated genomic and epigenetic features. This confirms the usefulness of lab systems in the study of retroviral integration targeting;
however the importance of subtle differences between virus and vector, and primary and transformed cell lines has yet to be explored.

It has been shown that a minority of the histone modification preferences of HIV-1 can be attributed directly to LEDGF-mediated targeting [6,17]. It is unclear whether components of the PIC interact with modified histones, with alternative host tethers that distinguish between histone modifications, or with indirectly correlated features marked by the preferred modifications to attain the observed distributions. Genomic loci that hosted insertions in more than one infection were particularly enriched for the favored modifications and poor in the disfavored modifications with two notable exceptions – trimethylated H3K79 and H4K20 are normally associated with heterochromatin and negatively correlated with integration site distributions. “Hot sites” in the genome were found in regions more enriched in these modifications. Interestingly, the H3K79me3 modification was more disfavored in a previously generated set of vector integration sites in Jurkat T cells (Fig 4-5C) [6]. H3K79me3 has been identified also as a marker of gene activation in CD4+ T cells [28]; therefore the observed relative preference for this marker could reflect a dominance of targeting to active genes over the disfavoring of heterochromatin. H4K20me3 is a marker of transcriptional repression in T cells [28]. Therefore, the enhanced targeting to regions rich in this modification may reflect previously unobserved co-localization of this modification with more attractive markers, or it may result from specific host-virus interactions.
We identified 960 genomic loci repeatedly targeted in separate infections for integration. As we might expect, these sites are found in extremely favorable global and local regions of the genome, but statistical analysis reveals that their attractiveness is not fully explained by the known preferences of HIV-1. Similarly there are several regions across the genome where we observe more integration sites than expected. These findings suggest that further work is necessary to identify the features that draw the PIC to these locations of the genome. Several host factors in addition to LEDGF have been found to associate with the PIC and IN [37-40] and their roles in targeting are unknown. Additionally, recent evidence presented in Chapters 2 and 3 suggest that different pathways through the nuclear pore significantly affect the chromatin available to the PIC [19,20]. Intriguingly, the chromatin binding protein EED was reported to bind and colocalize with HIV-1 IN near the nuclear pore [41]. It will be of interest to determine whether intra-nuclear arrangement of genomic regions with respect to the nuclear pore or nuclear transport factors correlate with preferences in targeting and add to models predicting integration intensity.

Acknowledgements

We would like to thank members of the Bushman lab for discussion and suggestions.
References


CHAPTER 5 – DYNAMIC REGULATION OF HIV-1 mRNA

POPULATIONS ANALYZED BY SINGLE MOLECULE ENRICHMENT

AND LONG READ SEQUENCING

The contents of this chapter have been submitted for publication as part of:
Ocwieja KE, Sherrill-Mix S, Mukherjee R, Custers-Allen R, David P, Brown M,
Wang S, Link DR, Olson J, Travers K, Schadt E, Bushman FD. Dynamic
regulation of HIV-1 mRNA populations analyzed by single molecule
enrichment and long read sequencing.

Abstract

Alternative RNA splicing greatly expands the repertoire of proteins encoded
by genomes. Next generation sequencing (NGS) is attractive for studying alternative
splicing because of the efficiency and low cost per base, but short reads typical of
NGS only report mRNA fragments containing one or few splice junctions. Here we
used single molecule amplification and long-read sequencing to study the HIV-1
provirus, which is only 9700 bp in length but encodes 9 major proteins via
alternative splicing. Our data showed that the clinical isolate HIV-1\textsubscript{896} produces at
least 109 different spliced RNAs, including a previously unappreciated \textasciitilde1kb class of
messages, two of which encode new proteins. HIV-1 message populations differed
between cell types, longitudinally during infection, and among T cells from different
human donors. These findings open a new window on a little-studied aspect of HIV-
1 replication, suggest therapeutic opportunities, and provide advanced tools for the
study of alternative splicing.
**Introduction**

Alternative splicing greatly expands the information content of genomes by producing multiple mRNAs from individual transcription units. Approximately 95% of human genes with multiple exons encode RNA transcripts that are alternatively spliced, and mutations that affect alternative splicing are associated with diseases ranging from cystic fibrosis to chronic lymphoproliferative leukemia [1-5]. Work to decipher an RNA “splicing code” has revealed that multiple interactions between trans-acting factors and RNA elements determine splicing patterns, though regulation is little understood for most genes [6].

The integrated HIV-1 provirus is ~9700 bp in length and has a single transcription start site, but according to published literature yields at least 47 different mRNAs encoding 9 proteins or polyproteins, making HIV an attractive model for studies of alternative splicing [7]. HIV mRNAs fall into three classes: the unspliced RNA genome, which encodes Gag/Gag-Pol; partially spliced transcripts, approximately 4kb in length, encoding Vif, Vpr, a one-exon version of Tat, and Env/Vpu; and completely spliced mRNAs of roughly 2kb encoding Tat, Rev and Nef. Additional rare “cryptic” splice donors (5’ splice sites) and acceptors (3’ splice sites) contribute even more mRNAs [8-13]. A complex array of positive and negative cis-acting elements surrounding each splice site regulates the relative abundance of the HIV-1 mRNAs, and disrupting the balance of message ratios impairs viral replication in several models [14-21]. Studies have suggested strain specific splicing patterns may exist [7,22]. However, detailed studies of complete message populations have
not been reported for clinical isolates of HIV-1.

Several groups have demonstrated tissue- and differentiation-specific splicing of cellular genes [2,23,24]. Importantly for HIV, these include changes during T cell activation, raising the question of how cell-specific splicing affects HIV replication [25,26]. While most studies of HIV-1 splicing have been conducted in cell lines using lab adapted viral strains, limited work in PBMCs from infected patients, monocytes and macrophages have suggested that differences may indeed exist in relevant cell types [9,22,27,28]. Moreover, human splicing patterns differ between individuals, but such polymorphisms have not been investigated in the context of HIV infection [29,30].

Here we use deep sequencing to comprehensively characterize the transcriptome of an early passage clinical isolate, HIV-1\textsubscript{89.6} [31], in primary CD4\textsuperscript{+} T cells from seven human donors and in the human osteosarcoma (HOS) cell line. Many deep sequencing techniques provide short reads, which rarely query more than a single exon-exon junction. To distinguish the full structure of HIV-1 mRNAs, which can contain several splice junctions, we used Pacific Biosciences (PacBio\textsuperscript{®}) sequencing technology, which yields read lengths up to 10kb [32]. We used RainDance Technologies single-molecule PCR enrichment to preserve ratios of RNAs during preparation of sequencing templates. We identified previously published and novel HIV-1 transcripts and determined that HIV-1\textsubscript{89.6} encodes a minimum of 109 different splice forms. These included a new size class of transcripts, some of which contain novel open reading frames (ORFs) that encode new proteins. We also
found significant variation between cell types, over time during infection of HOS cells, and among individuals. These data reveal unanticipated complexity and dynamics in HIV-1 message populations, begin to clarify a little-studied dimension of HIV-1 replication, and suggest possible targets for therapeutic interventions.

Methods

Cell culture and viral infections

HIV-189,6 was generated by the University of Pennsylvania Center for Aids Research. 293T cells were transfected with plasmid containing the viral clone, and harvested virus was passaged in SupT1 cells once.

Human osteosarcoma cells expressing CD4 and CCR5 (HOS-CD4-CCR5) were grown and passaged in DMEM with GlutaMAX (Invitrogen) supplemented with 10% FBS (Sigma-Aldrich), and 1 μg/mL Puromycin (Calbiochem). Low passage (<20) cells were plated in 10 cm plates and grown overnight to 80% confluency. HIV-189,6 equivalent to 2.25 micrograms of p24 was added to each plate in a total of 3mL D10. At 4hpi, cells were washed twice in PBS and media was replaced with 10mL fresh D10 with puromycin and 21.5 μM Nelfinavir to prevent second round infection. At 18, 24, and 48hpi, cells were washed in PBS, trypsinized, and pelleted.

Primary CD4+ T cells were isolated by the University of Pennsylvania Center for AIDS research Immunology core from apheresis product (Subject 1) or 40 mL whole blood (Subjects 2-7) using the RosetteSep Human CD4+ T Cell Enrichment Cocktail (StemCell Technologies). All donors were healthy and genotyped as below.
T Cells were stimulated for 3 days at 0.5 x10^6 cells per milliliter in R10 media (RPMI 1640 with GlutaMAX (Invitrogen) supplemented with 10% FBS (Sigma-Aldrich) with 100 units U/mL recombinant IL2 (Novartis) + 5ug/mL PHAL (Sigma-Aldrich). Cells from each donor were infected in triplicate. For each infection, three million cells were plated in each of 2 wells of a six well plate in 3 mL R10 media containing ~270 ng p24 of HIV-1^89.6 per 10^6 cells. Cells were infected by spinoculation for 2 hr at 1200-x g and 37°C. Plates were incubated at 37°C for an additional 2 hours, and then cells were pooled and volume increased to 12 mL R10 media with 100 U/mL IL2 and 100 pg/mL Enfuvirtide (T-20) to restrict infection to a single round. We have observed that ~5-6% of cells become infected at 48hpi using this protocol (data not shown). At 48hpi, cells were washed twice in PBS and pelleted for RNA preparation.

The following reagents were obtained through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH: HOS-CD4-CCR5 from Dr. Nathaniel Landau [33,34], Nelfinavir, and T-20, Fusion Inhibitor from AIDS, NIAID.

**Genotyping of the CCR5 allele**

Genomic DNA was isolated from T cells (subject 1) or from buccal swab (subjects 2-7) using DNeasy Blood and Tissue Kit (Qiagen). 35 ng gDNA and each of three control plasmid mixes containing CCR5 alleles (WT, Δ32, or both) were amplified in 25 uL with 200 pmols primers CCR5-F and CCR5-R (Table 5-1) using Go Taq Green Master Mix (Promega) as follows: 5 min at 95°C; 35 cycles of 30s at 95°C,
30s at 60°C, and 45s at 72°C; and finally 2 min at 72°C. Amplification products were resolved by 2% agarose gel electrophoresis and stained with ethidium bromide.

RNA and Reverse Transcription

Total cellular RNA was purified using the Illustra RNA kit (GE Life Sciences, Fairfield, CT) from 5x10^6 cells per infection. Viral cDNA was made using a reverse transcription primer complementary to a sequence in U3 (RTprime, Table 5-1 and Figure 5-1). We used Superscript III reverse transcriptase (Invitrogen) in the presence of RNaseOUT (Invitrogen) to conduct first strand cDNA synthesis from equal amounts of total cellular RNA from each HOS-CD4-CCR5 time point (15.2 μg) and from each T cell infection (3 μg) according to manufacturers instructions for gene specific priming of long cDNAs, and then treated with RNaseH (Invitrogen). We checked for full reverse transcription of the longest (unspliced) viral cDNAs by PCR using primers that bind in the first major intron of HIV-1 (keo003, keo004, Table 5-1 and Figure 5-1, data not shown)

Single molecule amplification

RainDance Technologies amplification was performed using a protocol similar to that reported [35]. PCR droplets were generated on the RDT 1000 using the manufacturer’s recommended protocol. The custom primer libraries for this study contained 18 (HOS-CD4-CCR5 cells) or 20 (primary T cells) PCR primer pairs designed to amplify different HIV RNA isoforms (Table 5-1).
Figure 5-1 Locations of DNA oligonucleotides used in this study with respect to genes and exons of the HIV-1<sub>env</sub> genome. Refer to Table 5-1 for descriptions. RT primer used to make viral cDNA is shown in white. Primer pairs are indicated by label and color. The two components of split primers are joined by dashed lines. Genes are labeled above the black bar with exons incorporated in their transcripts colored to match below.
Table 5-1 - DNA oligonucleotides used in this study

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence (5’→3’)*</th>
<th>Description</th>
<th>Coordinates</th>
<th>Split*</th>
<th>Cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCR5-F</td>
<td>TGGTGGCCGTGTGGTGCTTC</td>
<td>Forward primer for detecting CCR5 d32 mutation</td>
<td>9309:9333</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>CCR5-R</td>
<td>AGCGGCGAGCAGCAGCCCAAAG</td>
<td>Reverse primer for detecting CCR5 d32 mutation</td>
<td>939:961</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>RTPrime</td>
<td>CTCCACACTAATCCTGCTCTCCG</td>
<td>RT primer for making cDNA from HIV mRNA transcripts</td>
<td>1538:1559</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>keo003</td>
<td>TTAGAAACATCAGAGGCTGTAG</td>
<td>Forward primer to check for presence of full length (9kb) cDNA</td>
<td>5987:6004</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>keo004</td>
<td>TGGAAGATGTTGAGATTATGGTC</td>
<td>Reverse primer to check for presence of full length (9kb) cDNA</td>
<td>8458:8475</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F1.1.1</td>
<td>ACAGCGCAAGAGACTCC</td>
<td>Forward primer to determine ratio of spliced to incompletely spliced forms</td>
<td>550:569</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R1.1.1</td>
<td>TGCCAAGAATCCCGTTCAC</td>
<td>Reverse primer to determine ratio of spliced to incompletely spliced forms</td>
<td>5434:5456</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F1.1.2</td>
<td>AAAGTGTGTGGCCCCTCTGT</td>
<td>Forward primer to determine ratio of spliced to incompletely spliced forms</td>
<td>564:583</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R1.1.2</td>
<td>AGTGCTGTGGATATTTACACAAC</td>
<td>Reverse primer to determine ratio of spliced to incompletely spliced forms</td>
<td>5984:6002</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F1.2</td>
<td>AGTCTGTGGTGACTTCTGG</td>
<td>Forward primer to determine relative use of exons 2-5</td>
<td>549:567</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R1.2</td>
<td>CTGTCTGTGCTGCTTC</td>
<td>Reverse primer to determine relative use of exons 2-5</td>
<td>6353:6371</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F1.3</td>
<td>CAAGTAGTGTGTGGCGCCGTC</td>
<td>Forward primer to determine ratios of incompletely spliced (~4kb) transcripts</td>
<td>679:697</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R1.3</td>
<td>GTGGTAGTGTGGTCTCTTC</td>
<td>Reverse primer to determine ratios of incompletely spliced (~4kb) transcripts</td>
<td>6038:6043; 8368:8378</td>
<td>s HOS, T</td>
<td></td>
</tr>
<tr>
<td>F1.4</td>
<td>GATCTCTCGGGCTAGCAGACT</td>
<td>Forward primer to determine ratios of completely spliced (~2kb) transcripts</td>
<td>4929:4946</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R1.4</td>
<td>GGGAGAGGGTGGCTTT</td>
<td>Reverse primer to determine ratios of completely spliced (~2kb) transcripts including exon 2</td>
<td>4929:4946</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F1.5</td>
<td>TTTGGAAGGACCAGCAA</td>
<td>Forward primer to determine ratios of transcripts including exon 2</td>
<td>5994:6013</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R1.5</td>
<td>CTGTCTGTGAGGAGGTCTTC</td>
<td>Reverse primer to determine ratios of transcripts including exon 2</td>
<td>563:582</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F2.1.1</td>
<td>CCGTCTGTGTGTGCAGCTCTG</td>
<td>Forward primer to determine frequency of splicing from D1 to A1a</td>
<td>4568:4587</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R2.1.1</td>
<td>TTGCTGCATGTTGCTGTAG</td>
<td>Reverse primer to determine frequency of splicing from D1 to A1a</td>
<td>736:742; 4542:4551</td>
<td>s HOS, T</td>
<td></td>
</tr>
<tr>
<td>F2.1.2</td>
<td>GCGACTGCAGCAGATG</td>
<td>Forward primer to characterize splicing downstream when D1 is spliced to A1a</td>
<td>5984:6002</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R2.1.2</td>
<td>AGGTCTCTGCTCGTCTTC</td>
<td>Reverse primer to characterize splicing downstream when D1 is spliced to A1a</td>
<td>5995:6016</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>F2.1.2-1</td>
<td>GAGCTCGCTGTTGCTGAAG</td>
<td>Forward primer to characterize splicing upstream when D1a is spliced to A1</td>
<td>693:710</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R2.1.2</td>
<td>ATCTCTGTGCCTACTGCTC</td>
<td>Reverse primer to characterize splicing upstream when D1a is spliced to A1</td>
<td>4715:4720; 4912:4924</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F2.1.3.2</td>
<td>GACCGGAAAGCGAAAGAG</td>
<td>Forward primer to characterize splicing upstream when D1a is spliced to A3</td>
<td>650:667</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R2.1.3.2</td>
<td>CGACACCAATCTCTGTC</td>
<td>Reverse primer to characterize splicing upstream when D1a is spliced to A3</td>
<td>4715:4720; 5776:5787</td>
<td>s HOS, T</td>
<td></td>
</tr>
<tr>
<td>Primer</td>
<td>Sequence (5'→3'*</td>
<td>Description</td>
<td>Coordinates</td>
<td>Split*</td>
<td>Cell ***</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------</td>
<td>-------------</td>
<td>-------------</td>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td>F2.2.1</td>
<td>AGACAGCGGAGAAAGCTC</td>
<td>Forward primer to determine frequency of splicing from D4 to A5</td>
<td>5985:6003</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R2.2.1</td>
<td>GCTTGTTGGGTATAGAAAGAGC</td>
<td>Reverse primer to determine frequency of splicing from D4 to A5</td>
<td>6695:6717</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F2.2.3</td>
<td>CACAGCAAGAGGCGAG</td>
<td>Forward primer to determine upstream splicing events when D5 is spliced to A7</td>
<td>715:731</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R2.2.3</td>
<td>TGGGAGAGGCTTTATATTC</td>
<td>Reverse primer to determine upstream splicing events when D5 is spliced to A7</td>
<td>6722:6729; 8368:8379</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F2.3.1</td>
<td>AGAATTCGGCTTGCCTGAAA</td>
<td>Forward primer to determine upstream splicing events when D4 is spliced to A7a</td>
<td>692:709</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R2.3.1</td>
<td>CCGTGCTAATTTGCTTTTG</td>
<td>Reverse primer to determine upstream splicing events when D4 is spliced to A7a</td>
<td>6037:6043; 8334:8345</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F2.3.2</td>
<td>CAGGACTCGGCTTGGCTG</td>
<td>Forward primer to determine upstream splicing events when D4-A7b splicing occurs</td>
<td>691:707</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R2.3.2</td>
<td>GAATATCCCTGCTGCTTTG</td>
<td>Reverse primer to determine upstream splicing events when D4-A7b splicing occurs</td>
<td>6036:6043; 8340:8351</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F2.4.1</td>
<td>CAGGACAGAGACCTCC</td>
<td>Forward primer to determine relative use of splice acceptors in A7 and A8 clusters</td>
<td>5988:6004</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R2.4.1</td>
<td>TGTTAGGATGGTCTGAGCA</td>
<td>Reverse primer to determine relative use of splice acceptors in A7 and A8 clusters</td>
<td>9273:9291</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F2.4.2</td>
<td>ATCTCTCGAGGACAGCTC</td>
<td>Forward primer to determine upstream splicing events when D4 is spliced to A8</td>
<td>680:698</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R2.4.2</td>
<td>CCTGCTGTGTGTGTTGCTTTT</td>
<td>Reverse primer to determine upstream splicing events when D4 is spliced to A8</td>
<td>6037:6043; 9177:9190</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F3.1</td>
<td>AATCCCAAAGTTCAAGAGTAG</td>
<td>Forward primer for unbiased analysis of splicing, first third of genome</td>
<td>4658:4679</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R3.1</td>
<td>AGGGATCTTTCGCTCGTG</td>
<td>Reverse primer for unbiased analysis of splicing, first third of genome</td>
<td>5988:6005</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F3.2</td>
<td>GGGAAATGAGAGAAAGGAGA</td>
<td>Forward primer for unbiased analysis of splicing, second third of genome</td>
<td>6663:6681</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R3.2</td>
<td>GAGGAGAGGAGCAGGTTC</td>
<td>Reverse primer for unbiased analysis of splicing, second third of genome</td>
<td>8496:8514</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F3.3</td>
<td>CCGAAGAAGACAAAGAGGAAG</td>
<td>Forward primer for unbiased analysis of splicing, last third of genome</td>
<td>8402:8422</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R3.3</td>
<td>CTTGTCATCGACATCAGAAAAG</td>
<td>Reverse primer for unbiased analysis of splicing, last third of genome</td>
<td>9201:9220</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>F4.1</td>
<td>GTGTGCCCCTCTGTGTTG</td>
<td>Forward primer for full transcript structures</td>
<td>557:574</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>R4.1</td>
<td>CCATCCAGGTCGATGG</td>
<td>Reverse primer for full transcript structures</td>
<td>9194:9211</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>keo056</td>
<td>CAGGGACCGGAAACGGAAG</td>
<td>Forward primer to clone full length transcripts</td>
<td>646:666</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>keo057</td>
<td>GGAAAGTCCCAGGAGGAAGT</td>
<td>Reverse primer to clone full length transcripts</td>
<td>9429:9449</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>keo070</td>
<td>AGACTCTAGATTTGCCTCTCAAGCAG</td>
<td>Forward quikchange primer for plRES2-AcGF1-Tat1 to make 1ex tat</td>
<td>6016:6043; 8368:8386</td>
<td>HOS, T</td>
<td></td>
</tr>
<tr>
<td>keo071</td>
<td>TCGAGGCTGGAGAGCAATTTTCAGATGAGAAACTTTGTGACTG</td>
<td>Reverse quikchange primer for plRES2-AcGF1-Tat1 to make 1ex tat</td>
<td>6016:6043; 8368:8386</td>
<td>HOS, T</td>
<td></td>
</tr>
</tbody>
</table>

*For amplification primers used on samples from HOS cells, each primer had a fiveimer of 5'-GGTAT-3' appended to the 5' end to control for any biases in blunt end ligation step of PacBio SMRTBell hairpin loop attachment.

**Coordinates refer to HIV-189.6 reference sequence. Primers that span a known conserved or cryptic exon-exon junction are designated as split ("s") and the coordinates reflect the removed intron.

***Most primers were used on HIV RNA/cDNA from both HOS-CD4-CCR5 cells (HOS) and CD4-positive T cells (T). Primers that were used to amplify template from only one cell type are indicated.
Single molecule sequencing

Amplicon products from the RainDance PCR droplets were converted to SMRTbell™ templates using the PacBio® RS DNA Template Preparation Kit. Sequencing was performed by Pacific Biosciences using the PacBio SMRT™ sequencing technology as described [32]. Prior to sequence acquisition, hairpin adapters were ligated to each DNA template end so that DNA polymerase could traverse DNA molecules multiple times during rolling circle replication (SMRTbell™ template sequencing [36]), allowing error control by calculating the consensus ("circular consensus sequence" or CCS). Sequencing data was collected in 45 minute movies.

Bulk RT-PCR and cloning

For cloning, full length transcripts were amplified from cellular RNA using the Onestep RT-PCR kit (Qiagen) with primers keo056 and keo057 (Table 5-1) with the following amplification: 5 cycles of 30s at 94°C, 12s at 56°C, 40s at 72°C; then 30 cycles of 30s at 94°C, 14s at 56°C, 40s at 72°C; and finally 10 min at 72°C. For verification of dynamic changes primers F1.2 and R1.2 were used with 35 cycles of 30s at 94°C, 30s at 56°C, and 45s at 72°C followed by 10 min at 72°C. Products were resolved on agarose gels (Nusieve 3:1, Lonza for verification of dynamic changes, Invitrogen for cloning) stained with ethidium-bromide (Sigma) for visualization, or SYBR Safe DNA gel stain (Invitrogen) for cloning. DNA was purified using Qiaquick gel extraction kit (Qiagen) and cloned using the TOPO TA cloning kit (Invitrogen).
Plasmid DNA was prepared using Qiaprep Spin Miniprep kit (Qiagen). Inserts were identified and verified using Sanger sequencing.

The inserted cDNAs of transcripts with exon structures 1-4-8c (tat^8c), 1-4b-8c (ref), and 1-5-8c, as well as tat (2 exon), rev and nef cDNAs were excised from TOPO vector by restriction digest using EcoRI (New England Biolabs) and ligated into the expression vector pIRES2-AcGFP1 (Clontech). This vector contains a CMV promoter, and encodes and IRES and GFP following the inserted gene for production of a bicistronic message. Correct ligation was confirmed by Sanger sequencing. Plasmid encoding 1-exon Tat was created from the tat plasmid using the Quikchange II XL Site directed mutagenesis kit (Stratagene) to introduce a stop codon after the first exon with primers keo070 and keo071 (Table 5-1).

**Western Blotting**

HEK-293T cells were transfected as above with expression constructs in 6 well plates, 4 μg plasmid DNA per well. At 40 hours post transfection, 3 μL of DMSO or 10mM MG132 (EMD Chemicals) in DMSO was added per well. Cells were lysed after 3h MG132 or DMSO treatment in 150 mM NaCl, 50 mM tris-HCl pH 8.0, 2 mM EDTA, 1% NP-40, 0.5% deoxycholic acid, 0.1% SDS with protease inhibitors (Roche, 1697498). Protein was separated by SDS-PAGE using 4-15% Tris-HCl gradient gels (Biorad). Proteins were detected by immunoblotting using a mouse antibody that recognizes the carboxy terminus of HIV-1 Nef diluted 1:1000 in 5% milk (gift of Dr. James Hoxie) [37]. HRP conjugated secondary rabbit-anti-mouse antibody (p0260,
DAKO) was used for detection with SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific). Beta-tubulin was used as a loading control, detected by the HRP conjugated antibody (ab21058, Abcam).

**Tat activity assay**

TZM-bl cells (gift of Dr. Robert W. Doms) were grown in D10 medium (DMEM supplemented with 10% FBS (Sigma-Aldrich) in a 96 well plate to 80% confluency. These cells stably express a Tat-inducible luciferase. Cells were transiently transfected by lipofection using the Lipofectamine 2000 reagent (Invitrogen) according to the manufacturer’s instructions. Transfections were done in quadruplicate for each expression construct and the empty pIRES2-AcGFP1 vector with a total of 0.2 μg of DNA per well. Mock transfections were also performed. Expression of different plasmid constructs was confirmed to be equivalent by fluorescence microscopy, counting GFP-expressing cells (data not shown). Luciferase production was measured as emitted luminescence using the Bright-Glo luciferase assay system (Promega).

**Rev activity assay**

In order to determine whether the products expressed from our cDNAs had Rev activity, we co-transfected (as above) HEK-293T cells with each of our plasmids and pCMVGagPol-RRE-R, a reporter plasmid from which Gag and Pol are expressed in a Rev-dependent manner (gift of David Rekosh) [38]. Cells were plated in 12 well
plates and transfected with a total of 2.1 μg plasmid DNA per well: 0.7 μg pCMVGagPol-RRE-R plus 0.7 μg of experimental plasmid or a mix of 0.7 μg of the rev plasmid with 0.08 μg (Rev low), 0.23 μg (Ref medium) or 0.7 μg (Ref high) of the plasmid containing the ref cDNA insert. Empty vector (pIRES2-AcGFP) was added to bring the total DNA to 2.1 μg. Reporter gag expression was measured by flow cytometry and by ELISA of p24 (Capsid) in supernatants (below).

**HIV Replication assays**

In order to determine whether expression of any of the cloned cDNAs affected HIV-1\textsubscript{89,6} infection efficiency, we transiently transfected each of the expression plasmids into HOS-CD4-CCR5 cells by lipofection essentially as above in 24 well plates with 0.8 μg plasmid per well. Cells were infected with HIV-1\textsubscript{89,6} equivalent to 125 ng p24 in 300 μL D10 medium per well 24 hours after transfection. Spinoculation was performed at 1200g for 2hr at 37°C. Cells were then incubated at 37°C for an additional 4 hours, after which they were washed with PBS and fresh media was applied. Cells were analyzed for intracellular Gag expression by flow cytometry (below) and for release of p24 into supernatant by ELISA by the University of Pennsylvania Center for AIDS Research.

**Flow cytometry**

1-2x10^6 cells were stained per co-transfection for flow cytometry. All incubations were at room temperature. Cells were first washed in PBS and then
twice in FACS wash buffer (PBS, 2.5% FBS, 2mM EDTA). Cells were fixed and permeabilized with CytoFix/CytoPerm (BD) for 20 minutes and washed with Perm-Wash Buffer (BD) before staining with anti-HIV-Gag-PE (Beckman Coulter) for 60 min. Finally cells were washed in FACS wash buffer and resuspended in 3% PFA. Samples were run on a LSRII (BD) and analyzed with FlowJo 8.8.6 (Treestar). Cells were gated as follows: lymphocytes (SSC-A by FSC-A), then singlets (FSC-A by FSC-H), then GFP+ cells (FSC-A by GFP), then by Gag expression (FSC-A by Gag).

Data Analysis

Raw sequence reads with multiple read-throughs were processed to produce circular consensus sequences (CCS). Raw reads were also retained to help in primer identification and to avoid biasing against long reads. Reads were aligned against the human genome using Blat [39]. Mis-primed reads matching the RT primer, reads with a CCS length shorter than 40 nucleotides or raw length shorter than 100 nucleotides and reads matching the human genome were discarded. Filtered reads were aligned against the HIV-189,6 reference genome. Potential novel donors and acceptors were found by filtering putative splice junctions in the Blat hits for a perfect sequence match 20 bases up and downstream of the junction, ignoring homopolymer errors, and requiring that one end of the junction be a known splice site. Local maximums within a 5-nt span with >9 such junctions were called as novel splice sites.

Filter-passed reads were aligned against all expected fragments based on
primers and known and novel junctions. Primers were identified in CCS reads by an edit distance ≤1 from the primer in the start or end of the read, in raw reads by an edit distance ≤5 from a concatenation of the primer, hairpin adapter, and the reverse complement of the primer, and in both types of reads by a Blat hit spanning an entire expected fragment.

Gaps in Blat hits were ignored if ≤10 bases long or in regions of likely poor read quality ≤20 bases long where an inferred insertion of unmatched bases in the read occurred at the same location as skipped bases in the reference. Any Blat hits with a gap >10nt remaining in the query read were discarded. If HIV sequence was repeated in a given read (likely due to PacBio® circular sequencing), the alignments were collapsed into the union of the coverage. Gaps in the HIV sequence found in uninterrupted query sequence were called as tentative introns. Splice junctions were assigned to conserved or previously identified (published or in this work) splice sites and reads appearing to contain donors or acceptors further than 5 nucleotides away from these sites were discarded. Reads with Blat hits outside the expected primer range were discarded from that primer grouping. The assigned primer pair, observed junctions and exonic sequence were used to assign each read to a given splice-form (specific transcript structure) or set of possible splice-forms. Partial sequences that did not extend through both primers were assigned to specific transcripts if the read contained enough information to rule out all other splice-forms or if all other possible splice-forms contained rare (<1% usage) donors or acceptors. Otherwise, the read was called indeterminate. Note that the above
computational data analysis was performed by Scott Sherrill-Mix.

To calculate the ratios of transcripts within the partially spliced class, we counted the number of reads for each assigned splice-form amplified by primer pair 1.3 and divided by the total number of assigned partially spliced reads amplified with these primers (Table 5-1 and Fig. 5-1). Assigned sequences amplified with primer pairs 1.4 and 4.1 (full length cDNAs, T cells only) were used to calculate ratios of transcripts within each of the two completely splice classes (~2kb and ~1kb). To compare ratios of ~2kb transcripts calculated within reads from primer pairs 1.4 and 4.1, we normalized ratios from pair 4.1 to the nef 2 transcript (containing exons 1, 5, and 7). Due to size biases inherent in the approach, we did not compare across size classes, and unspliced transcripts were not included in ratio analysis. For all ratio analysis, transcripts including cryptic or novel junctions were counted only if they appeared in at least 5 reads; otherwise they were excluded from the analysis and from the count of total assigned reads.

To estimate the minimum total number of transcripts present, partial sequence reads were included. Each exon-exon junction occurring in at least 5 reads and not previously assigned to one of the 78 fully-characterized transcripts (Fig. 2) was counted as evidence of an additional transcript (47 additional junctions were detected). If two such junctions could conceivably occur in a single mRNA, we counted only one unless we could verify from sequence reads that they were amplified from separate cDNAs, yielding a minimum of 31 additional transcripts.
For studies of transcript dynamics, reads from primer pairs 1.2, 1.3 and 1.4 containing junctions between D1 or any donor and each of five mutually exclusive acceptors, A3, A4c, A4a, A4b, A5, and A5a, were collected and their ratios calculated. Statistical modeling was performed by Scott Sherrill-Mix. The relationships between read counts of mutually exclusive acceptors and (a) infection time, (b) subject and (c) cell type were investigated using generalized linear models assuming a Poisson distribution of error with an extra quasi-Poisson variance term to allow for overdispersion in the data. We assumed a linear effect of time. Differences in sequencing effort were accounted for by including the total number of mutually exclusive reads as a fixed offset. All alignment and statistical analyses were performed in R 2.14.0 [R Development Core 40].

**Data Access**

Sequence data will be made available in the SRA database upon acceptance of the above mentioned manuscript for publication. Data is also stored at: [http://microb230.med.upenn.edu/Downloads/owieja_et_al_2012_data.zip](http://microb230.med.upenn.edu/Downloads/owieja_et_al_2012_data.zip).

Consensus sequence for HIV-1<sub>896</sub> is at the above link.

**Results**

**Sequencing HIV-1 transcripts produced in primary T cells and HOS cells**

In order to characterize HIV-1 transcript populations, we prepared viral cDNA from primary CD4<sup>+</sup> T cells of seven different healthy human donors infected in
vitro with HIV-1<sub>89.6</sub>, an early passage dual-tropic clade-B clinical isolate (see methods, human donor data in Table 5-2) [31]. We also studied HIV messages produced in infected HOS cells engineered to express CD4 and CCR5 (HOS-CD4-CCR5) because these cells support efficient HIV replication and engineered variants are widely used in HIV research. HOS cells were harvested at 18, 24, and 48 hours post infection (hpi) to investigate longitudinal changes during infection, and for comparison to 48 hour-infected T cells.

**Table 5-2 - Human T cell donor information**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Donated Material</th>
<th>Sex</th>
<th>Ethnicity</th>
<th>CCR5 genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Apheresis product</td>
<td>Female</td>
<td>Data not collected</td>
<td>WT/WT</td>
</tr>
<tr>
<td>2</td>
<td>Whole Blood</td>
<td>Male</td>
<td>Caucasian</td>
<td>WT/WT</td>
</tr>
<tr>
<td>3</td>
<td>Whole Blood</td>
<td>Male</td>
<td>Caucasian</td>
<td>WT/WT</td>
</tr>
<tr>
<td>4</td>
<td>Whole Blood</td>
<td>Male</td>
<td>South Asian</td>
<td>WT/WT</td>
</tr>
<tr>
<td>5</td>
<td>Whole Blood</td>
<td>Male</td>
<td>Caucasian</td>
<td>WT/WT</td>
</tr>
<tr>
<td>6</td>
<td>Whole Blood</td>
<td>Female</td>
<td>Caucasian</td>
<td>WT/WT</td>
</tr>
<tr>
<td>7</td>
<td>Whole Blood</td>
<td>Male</td>
<td>East Asian</td>
<td>WT/WT</td>
</tr>
</tbody>
</table>

To preserve the relative proportions of template molecules while amplifying the cDNA, we used RainDance Technologies’ single molecule micro-droplet based PCR [35]. Droplet libraries containing multiple overlapping primer pairs were designed to query all message forms and allow later calculation of relative abundance (Table 5-1 and Fig. 5-1). Each primer was unique so that sequences could be assigned to a specific primer pair, which helped reconstruct the origin of sequence reads, and deduce message structures. Amplified DNA products were sequenced using Single Molecule Real-Time (SMRT®) technology from Pacific Biosciences [32,36]. We obtained 847,492 filtered reads of amplified HIV-1
transcripts in primary CD4+ T cells and 89,350 in HOS cells. The longest sequenced continuous stretch of HIV-1 cDNA was 2629 bp.

**Splice Donors and Acceptors**

The analysis presented below was done in collaboration with Scott Sherrill-Mix, a graduate student in the Bushman Lab. Alignments and computational work requiring programming as well as statistical analysis was done by Scott, while calculations of ratios and of minimal represented transcript forms were performed manually. Scott aligned PacBio® reads containing HIV sequences to the HIV-189.6 genome and identified candidate introns as recurring gaps in our sequences. Using this approach we observed splicing at each of the widely conserved major splice donors and acceptors and several published cryptic sites (Fig. 5-2, hereafter referred to by their identifications shown in this figure, “D” for donors, “A” for acceptors).

In addition, we identified 13 putative novel splice sites: 2 donors and 11 acceptors (Fig. 5-2 and Table 5-3). In order to be selected as a bona fide splice site, we required that the new acceptor or donor was observed spliced to previously reported splice donors or acceptors in >10 sequence reads in CD4+ T cells. Most of the new splice sites adhered to consensus sequences for the standard spliceosome (Table 5-3). However, there was one new donor upstream of D1 with a cytidine in place of the usual uracil 2 nucleotides downstream of the splice site. Similar “GC donors” appear in 1% of known splice junctions in humans [41]. Of the novel splice
Figure 5-2 - Mapping the splice donors and acceptors of HIV-189,6. PacBio® sequence reads of HIV-189,6 cDNA from infected HOS-CD4-CCR5 (HOS) and T cells were aligned to the HIV-189,6 genome shown in (A). Exons of conserved transcripts are colored according to the encoded gene, 2 and 3 are variably included. Conserved (black) and cryptic (brown) splice donors (D) and acceptors (A) are shown, numbered according to [7]. Gaps in HIV-1 sequence alignments with at least one end located at a published or verified splice donor or acceptor were defined as introns. For each base of the genome, the number of sequence reads in which that base occurred at the 5’ end (B) or 3’ end (C) of an intron is plotted by cell type. Putative novel splice donors (blue) and acceptors (red), numbered according to nearest published site, were defined as loci that were found in at least 10 reads to be at the 5’ and 3’ ends of introns respectively, in sequence alignments from T cell infections. Regions containing splice sites are enlarged for clarity. Asterisks indicate putative donors and acceptors adjacent to dinucleotides other than the consensus GT and AG. Analyzed with Scott Sherrill-Mix. (B) and (C) plotted by Scott.
Table 5-3 – HIV-1<sub>89.6</sub> splice donors and acceptors.

<table>
<thead>
<tr>
<th>ID</th>
<th>Coordinate*</th>
<th>Type</th>
<th>Usage Proportion**</th>
<th>Sequence***</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Splice Donors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1b</td>
<td>725</td>
<td>Novel</td>
<td>0.0102</td>
<td>ag</td>
</tr>
<tr>
<td>D1</td>
<td>742</td>
<td>Conserved</td>
<td>0.9389</td>
<td>tg</td>
</tr>
<tr>
<td>D1c</td>
<td>746</td>
<td>Cryptic</td>
<td>0.047</td>
<td>ga</td>
</tr>
<tr>
<td>D1a</td>
<td>4720</td>
<td>Cryptic</td>
<td>0.0001</td>
<td>ag</td>
</tr>
<tr>
<td>D2</td>
<td>4961</td>
<td>Conserved</td>
<td>0.0587</td>
<td>ag</td>
</tr>
<tr>
<td>D3</td>
<td>5462</td>
<td>Conserved</td>
<td>0.0686</td>
<td>ag</td>
</tr>
<tr>
<td>D4</td>
<td>6043</td>
<td>Conserved</td>
<td>0.824</td>
<td>ca</td>
</tr>
<tr>
<td>D5</td>
<td>6729</td>
<td>Cryptic</td>
<td>0.0004</td>
<td>at</td>
</tr>
<tr>
<td>D6</td>
<td>8422</td>
<td>Novel</td>
<td>0.0011</td>
<td>ag</td>
</tr>
<tr>
<td>A1a</td>
<td>4542</td>
<td>Cryptic</td>
<td>0.0001</td>
<td>ttcctaaaatAG</td>
</tr>
<tr>
<td>A1</td>
<td>4912</td>
<td>Conserved</td>
<td>0.0577</td>
<td>cgggtttattacAG</td>
</tr>
<tr>
<td>A2</td>
<td>5389</td>
<td>Conserved</td>
<td>0.2031</td>
<td>tgatgttatttcAG</td>
</tr>
<tr>
<td>A3</td>
<td>5776</td>
<td>Conserved</td>
<td>0.0396</td>
<td>ttattctatttcAG</td>
</tr>
<tr>
<td>A4c</td>
<td>5935</td>
<td>Conserved</td>
<td>0.1212</td>
<td>cttctaatggcaAG</td>
</tr>
<tr>
<td>A4a</td>
<td>5953</td>
<td>Conserved</td>
<td>0.0638</td>
<td>ttcataacaaaAG</td>
</tr>
<tr>
<td>A4b</td>
<td>5959</td>
<td>Conserved</td>
<td>0.089</td>
<td>aacaaagcattAG</td>
</tr>
<tr>
<td>A5</td>
<td>5975</td>
<td>Conserved</td>
<td>0.6139</td>
<td>atctctcattggcAG</td>
</tr>
<tr>
<td>A5a</td>
<td>5979</td>
<td>Novel</td>
<td>0.0964</td>
<td>cctatgccagaAG</td>
</tr>
<tr>
<td>A5b</td>
<td>5982</td>
<td>Novel</td>
<td>0.0077</td>
<td>atggcaggaagaAG</td>
</tr>
<tr>
<td>A6</td>
<td>6602</td>
<td>Cryptic</td>
<td>0</td>
<td>ccacactgtgattAC</td>
</tr>
<tr>
<td>A6a</td>
<td>6654</td>
<td>Novel</td>
<td>0.0003</td>
<td>actaatccccactAG</td>
</tr>
<tr>
<td>A7a</td>
<td>8334</td>
<td>Cryptic</td>
<td>0.0016</td>
<td>tctatagtaaatAG</td>
</tr>
<tr>
<td>A7b</td>
<td>8340</td>
<td>Cryptic</td>
<td>0.0114</td>
<td>gtaaatagagttAG</td>
</tr>
<tr>
<td>A7c</td>
<td>8344</td>
<td>Novel</td>
<td>0.0023</td>
<td>atagagttaggcAG</td>
</tr>
<tr>
<td>A7d</td>
<td>8362</td>
<td>Novel</td>
<td>0.0146</td>
<td>attcaccattatCG</td>
</tr>
<tr>
<td>A7</td>
<td>8368</td>
<td>Conserved</td>
<td>0.8738</td>
<td>cattatgcttttcAG</td>
</tr>
<tr>
<td>A7e</td>
<td>8381</td>
<td>Novel</td>
<td>0.02</td>
<td>gacccctccctccAG</td>
</tr>
<tr>
<td>A7f</td>
<td>8485</td>
<td>Novel</td>
<td>0.0007</td>
<td>tgggactttttttcG</td>
</tr>
<tr>
<td>A8d</td>
<td>9128</td>
<td>Novel</td>
<td>0.0137</td>
<td>tcttgattgatG</td>
</tr>
<tr>
<td>A8a</td>
<td>9144</td>
<td>Cryptic</td>
<td>0.0371</td>
<td>ctaccacacacaAG</td>
</tr>
<tr>
<td>A8c</td>
<td>9156</td>
<td>Novel</td>
<td>0.4571</td>
<td>aggctttttccccAG</td>
</tr>
<tr>
<td>A8</td>
<td>9164</td>
<td>Cryptic</td>
<td>0.0293</td>
<td>tcccgactttggcAG</td>
</tr>
<tr>
<td>A8b</td>
<td>9177</td>
<td>Cryptic</td>
<td>0.0048</td>
<td>gaactacacaccAG</td>
</tr>
<tr>
<td>A8e</td>
<td>9183</td>
<td>Novel</td>
<td>0.0036</td>
<td>cacaccacacaccAG</td>
</tr>
<tr>
<td>A8f</td>
<td>9221</td>
<td>Novel</td>
<td>0.013</td>
<td>gatggtgctacaAG</td>
</tr>
</tbody>
</table>

* Coordinate indicates last base of exon for splice donors, first base of exon for acceptors

**Usage proportion is calculated as the fraction of sequence reads spanning the indicated splice site that contain a splice event involving that site.

***Exon/intron boundaries are indicated by "|". +1 and +2 positions within the introns following splice donors (positions of consensus GT) are shown in capitals. Similarly -2 and -1 positions within the introns preceding acceptors (positions of consensus AG) are shown in capitals. Consensus for human splice donors is AG|GTRAGT and for splice acceptors, YTTYYYYYYNCAG|G [43].

146
acceptors, three were preceded by dinucleotides other than the consensus AG. Alternative dinucleotides are not known to support splicing at acceptors; however, it is possible our methods allowed us to observe rare aberrant splicing [42,43].

The majority of novel splice sites were identified in clusters near previously observed sites. Two infrequently used splice donors were detected near D1. Use of these in place of D1 would modify the 5' untranslated leader on all spliced mRNAs without affecting any known HIV-1 ORFs. One was novel and use of the other, 4 nucleotides downstream of D1, had only been observed in viruses in which D1 was mutated [7]. Similarly, two novel acceptors downstream of exon 5 are predicted to shorten exon 5 but should not affect the nef, env, or vpu ORFs encoded by mRNAs that use exon 5. Novel acceptors near A7 were not redundant. Two (A7c and A7d, Table 5-3) would add novel amino acids between the protein domains encoded by the first and second coding exons of tat and rev. Splicing at A7e is predicted to produce Tat-Rev and Rev-Env fusions, while use of A7f would create Tat-Env and Rev-Tat fusions. A final cluster of acceptors within the 3’ LTR included the most frequently used novel splice site, an acceptor we have termed A8c because it lies near A8, A8a, and A8b (discussed in detail below). While we have verified the use of A8c, other sites will require further study. Junctions in which splice sites were found are summarized in Table 5-4.

**Structures of spliced HIV-1Δ89.6 RNAs**

To quantify the populations of HIV-1 transcripts, Scott aligned all reads to the collection of 47 well-established spliced HIV-1 transcripts and detected 45 of them.
Table 5-4 – HIV-189.6 splice junction counts in sequences from T cells

<table>
<thead>
<tr>
<th>Splice Acceptors</th>
<th>Splice Donors</th>
<th>D1b*</th>
<th>D1c</th>
<th>D1a</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>D5</th>
<th>D6*</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>151</td>
<td>12738</td>
<td>113</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A1a</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A2</td>
<td>407</td>
<td>43287</td>
<td>138</td>
<td>3</td>
<td>2367</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A3</td>
<td>109</td>
<td>7241</td>
<td>23</td>
<td>6</td>
<td>1761</td>
<td>456</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A4a</td>
<td>150</td>
<td>10696</td>
<td>52</td>
<td>0</td>
<td>228</td>
<td>354</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A4b</td>
<td>257</td>
<td>15442</td>
<td>66</td>
<td>0</td>
<td>777</td>
<td>873</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A4c</td>
<td>255</td>
<td>20605</td>
<td>64</td>
<td>7</td>
<td>588</td>
<td>1254</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A5</td>
<td>1058</td>
<td>104005</td>
<td>356</td>
<td>2</td>
<td>7166</td>
<td>9209</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A5a*</td>
<td>210</td>
<td>7883</td>
<td>88</td>
<td>5</td>
<td>336</td>
<td>425</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A5b*</td>
<td>193</td>
<td>1301</td>
<td>36</td>
<td>1</td>
<td>71</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A6a*</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0</td>
</tr>
<tr>
<td>A7a</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>166</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>A7b</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1216</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A7c</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>252</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A7d</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>295</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A7</td>
<td>38</td>
<td>695</td>
<td>6</td>
<td>0</td>
<td>19</td>
<td>13</td>
<td>105508</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>A7e*</td>
<td>16</td>
<td>62</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>1763</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A7f</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>A8d*</td>
<td>0</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>159</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>A8a</td>
<td>1</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>411</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>A8c</td>
<td>13</td>
<td>287</td>
<td>4</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td>5142</td>
<td>1</td>
<td>49</td>
</tr>
<tr>
<td>A8</td>
<td>5</td>
<td>21</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>492</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>A8b</td>
<td>3</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>62</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A8e</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>45</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A8f</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>92</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Splice acceptors and splice donors are labeled as identified in Table 5-3. Widely conserved major splice acceptors and donors are in bold. Asterisks indicate novel splice and donors identified in this work. Displayed counts are sequence reads with gaps in the alignment to the HIV-189.6 consensus (introns) beginning at the indicated splice acceptor and ending at the indicated donor. For this analysis all filtered sequence reads from T cells were pooled.

(Fig. 5-3). Scott additionally aligned reads to the HIV-189.6 genome allowing all possible combinations of splice junctions – canonical, cryptic, or novel – determined from the sequencing data (above, Table 5-3). We observed an additional 32 complete transcripts, 19 of which were novel, and each of which was observed >5
times in the T cell data and in >2 human donors. The data also provides evidence for more novel splice junctions but in incomplete sequences, implying the existence of additional new transcripts (Table 5-4 and below). The full data set taken together provides evidence for at least 109 different HIV-1_89.6 transcripts in primary T cells (calculated using Table 5-4 and data not shown as described in methods).

Amplification primers that isolated the two main classes of spliced messages allowed us to determine the ratios of mRNAs in each (Fig. 5-3 and Table 5-5). Within the partially spliced class of transcripts env/vpu, tat (1-exon), vpr, and vif messages existed in an average ratio of 96:4:<1:<1 in CD4+ T cells. The ratio of nef:rev:tat:vpr within the ~2kb transcript class was 64:33:3:<1. Consistent with previous reports, the most abundant transcript in each class contained the splice junction from D1 to A5 (D1^A5) – an env/vpu transcript contributing 64% of the partially spliced class, and a completely spliced nef transcript contributing 47% of ~2kb messages (Fig. 5-3) [7,44]. The relatively low abundance of transcripts encoding Tat suggests that Tat sufficiently stimulates HIV transcription elongation at low concentrations, or that the tat transcripts must be efficiently translated. The ratios were confirmed using overlapping sequence reads obtained with alternate primer pairs and by endpoint RT-PCR analysis of HIV-1 RNAs (data not shown).

A Novel ~1kb Class of Completely Spliced Transcripts

Primers placed near the 5’ and 3’ ends of the HIV-1_89.6 genome amplified a second class of completely spliced transcripts ~1kb in length. In place of A7, these
**Figure 5-3 - Spliced transcripts produced from HIV-189.6.** HIV-189.6 transcripts in T cells for which the full message structure was determined are shown arranged by size class (unspliced genome, partially spliced or 4kb, completely spliced or 2kb, and a new completely spliced 1kb class). Thick bars correspond to exons, thin lines to excised introns. For the well conserved transcripts, encoded proteins are indicated. The relative abundance of each transcript within its size class is indicated by color according to the scale displayed. Asterisks denote transcripts that have not been reported previously to our knowledge. Of the 47 conserved HIV-1 transcripts, 3 were detected in fewer than 5 reads (one tat and two env/vpu messages, indicated, 0), and two messages were not detected and are not shown (one encoding Vpr and one encoding Env/Vpu). Depicted transcripts using novel or cryptic splice sites were each detected in at least five independent sequence reads across samples from at least 2 different human T cell donors. Calculations were performed by Karen Ocwieja, data plotted by Scott Sherrill-Mix.
<table>
<thead>
<tr>
<th>Class</th>
<th>Primer Pair*</th>
<th>Transcript number**</th>
<th>splicing</th>
<th>percent of class</th>
<th>&lt;5 reads****</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T Cells (48hpi)</td>
<td>HOS (18hpi)</td>
</tr>
<tr>
<td>1.3</td>
<td>Vif 2</td>
<td>D1*A1</td>
<td>0.21%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>Vpr 3</td>
<td>D1*A2</td>
<td>0.46%</td>
<td>0.00%</td>
<td>0.58%</td>
</tr>
<tr>
<td>1.3</td>
<td>Vpr 4</td>
<td>D1<em>A1[D2</em>A2]</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>Tat 5</td>
<td>D1*A3</td>
<td>1.92%</td>
<td>5.41%</td>
<td>6.43%</td>
</tr>
<tr>
<td>1.3</td>
<td>Tat 6</td>
<td>D1<em>A1[D2</em>A3]</td>
<td>0.62%</td>
<td>2.70%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>Tat 7</td>
<td>D1<em>A2[D3</em>A3]</td>
<td>0.14%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>Tat 8</td>
<td>D1<em>A1[D2</em>A2][D3*A3]</td>
<td>0.04%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 4</td>
<td>D1*A4c</td>
<td>9.56%</td>
<td>2.70%</td>
<td>3.51%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 8</td>
<td>D1<em>A1[D2</em>A4c]</td>
<td>0.27%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 12</td>
<td>D1<em>A2[D3</em>A4c]</td>
<td>0.36%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 16</td>
<td>D1<em>A1[D2</em>A2][D3*A4c]</td>
<td>0.01%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 3</td>
<td>D1*A4a</td>
<td>5.60%</td>
<td>5.41%</td>
<td>4.09%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 7</td>
<td>D1<em>A1[D2</em>A4a]</td>
<td>0.09%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 11</td>
<td>D1<em>A2[D3</em>A4a]</td>
<td>0.17%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 15</td>
<td>D1<em>A1[D2</em>A2][D3*A4a]</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 2</td>
<td>D1*A4b</td>
<td>8.72%</td>
<td>8.11%</td>
<td>8.19%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 6</td>
<td>D1<em>A1[D2</em>A4b]</td>
<td>0.58%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 10</td>
<td>D1<em>A2[D3</em>A4b]</td>
<td>0.31%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 14</td>
<td>D1<em>A1[D2</em>A2][D3*A4b]</td>
<td>0.03%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 1</td>
<td>D1*A5</td>
<td>63.64%</td>
<td>70.27%</td>
<td>74.27%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 5</td>
<td>D1<em>A1[D2</em>A5]</td>
<td>2.66%</td>
<td>2.70%</td>
<td>1.75%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 9</td>
<td>D1<em>A2[D3</em>A5]</td>
<td>4.50%</td>
<td>2.70%</td>
<td>1.17%</td>
</tr>
<tr>
<td>1.3</td>
<td>EnvVpu 13</td>
<td>D1<em>A1[D2</em>A2][D3*A5]</td>
<td>0.11%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.4</td>
<td>Vpr 1</td>
<td>D1<em>A2[D4</em>A7]</td>
<td>0.14%</td>
<td>0.00%</td>
<td>0.13%</td>
</tr>
<tr>
<td>1.4</td>
<td>Vpr 2</td>
<td>D1<em>A1[D2</em>A2][D4*A7]</td>
<td>0.01%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.4</td>
<td>Tat 1</td>
<td>D1<em>A3[D4</em>A7]</td>
<td>2.69%</td>
<td>5.00%</td>
<td>5.45%</td>
</tr>
<tr>
<td>1.4</td>
<td>Tat 2</td>
<td>D1<em>A1[D2</em>A3][D4*A7]</td>
<td>0.67%</td>
<td>0.43%</td>
<td>0.85%</td>
</tr>
<tr>
<td>1.4</td>
<td>Tat 3</td>
<td>D1<em>A2[D3</em>A3][D4*A7]</td>
<td>0.14%</td>
<td>0.65%</td>
<td>0.09%</td>
</tr>
<tr>
<td>1.4</td>
<td>Tat 4</td>
<td>D1<em>A1[D2</em>A2][D3<em>A3][D4</em>A7]</td>
<td>0.01%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.4</td>
<td>Rev 3</td>
<td>D1<em>A4c[D4</em>A7]</td>
<td>12.76%</td>
<td>1.30%</td>
<td>2.64%</td>
</tr>
<tr>
<td>1.4</td>
<td>Rev 6</td>
<td>D1<em>A1[D2</em>A4c][D4*A7]</td>
<td>0.28%</td>
<td>0.43%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.4</td>
<td>Rev 9</td>
<td>D1<em>A2[D3</em>A4c][D4*A7]</td>
<td>0.54%</td>
<td>0.00%</td>
<td>0.09%</td>
</tr>
<tr>
<td>1.4</td>
<td>Rev 12</td>
<td>D1<em>A1[D2</em>A2][D3<em>A4c][D4</em>A7]</td>
<td>0.03%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.4</td>
<td>Rev 2</td>
<td>D1<em>A4a[D4</em>A7]</td>
<td>7.18%</td>
<td>4.35%</td>
<td>3.83%</td>
</tr>
<tr>
<td>1.4</td>
<td>Rev 5</td>
<td>D1<em>A1[D2</em>A4a][D4*A7]</td>
<td>0.14%</td>
<td>0.00%</td>
<td>0.04%</td>
</tr>
<tr>
<td>1.4</td>
<td>Rev 8</td>
<td>D1<em>A2[D3</em>A4a][D4*A7]</td>
<td>0.32%</td>
<td>0.43%</td>
<td>0.13%</td>
</tr>
<tr>
<td>1.4</td>
<td>Rev 11</td>
<td>D1<em>A1[D2</em>A2][D3<em>A4a][D4</em>A7]</td>
<td>0.01%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.4</td>
<td>Rev 1</td>
<td>D1<em>A4b[D4</em>A7]</td>
<td>9.38%</td>
<td>8.91%</td>
<td>6.64%</td>
</tr>
<tr>
<td>1.4</td>
<td>Rev 4</td>
<td>D1<em>A1[D2</em>A4b][D4*A7]</td>
<td>0.56%</td>
<td>0.00%</td>
<td>0.09%</td>
</tr>
<tr>
<td>1.4</td>
<td>Rev 7</td>
<td>D1<em>A2[D3</em>A4b][D4*A7]</td>
<td>0.62%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1.4</td>
<td>Rev 10</td>
<td>D1<em>A1[D2</em>A2][D3<em>A4b][D4</em>A7]</td>
<td>0.07%</td>
<td>0.00%</td>
<td>0.04%</td>
</tr>
<tr>
<td>1.4</td>
<td>Nef 2</td>
<td>D1<em>A5[D4</em>A7]</td>
<td>46.86%</td>
<td>72.39%</td>
<td>71.96%</td>
</tr>
</tbody>
</table>

1.4 and Nef 2 | D1*A5[D4*A7] | 46.86% | 72.39% | 71.96% | 78.29% |

1.4 | Nef 3 | D1*A1[D2*A5][D4*A7] | 4.69% | 3.26% | 4.43% | 1.07% |

1.4 | Nef 4 | D1*A2[D3*A5][D4*A7] | 4.96% | 2.39% | 3.11% | 0.42% |

1.4 | Nef 5 | D1*A1[D2*A2][D3*A5][D4*A7] | 0.34% | 0.43% | 0.43% | 0.03% |

1.4 | Novel (Nef 9) | D1*A5a[D4*A7] | 3.57% |

1.4 | Novel (Nef 11) | D1*A2[D3*A5a][D4*A7] | 0.28% |

1.4 | Nef 1 | D1-A7 | 2.01% |

1.4 | Novel | D1*A5[D4*A7b] | 1.33% |

1.4 | Novel | D1*A5[D4*A7e] | 0.44% |
<table>
<thead>
<tr>
<th>Class</th>
<th>Primer Pair*</th>
<th>Transcript number**</th>
<th>Splicing</th>
<th>percent of class</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A5</td>
<td>D4^A8d</td>
<td>0.80%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A8d</td>
<td>0.32%</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A4c</td>
<td>D4^A8a</td>
<td>0.78%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A4b</td>
<td>D4^A8a</td>
<td>0.96%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A5</td>
<td>D4^A8a</td>
<td>2.09%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A1</td>
<td>D2^A5</td>
<td>D4^A8a</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A5a</td>
<td>D4^A8a</td>
<td>0.15%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A8a</td>
<td>1.90%</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Novel (Tat^8c 1)</td>
<td>D1^A3</td>
<td>D4^A8c</td>
<td>0.64%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel (Ref 3)</td>
<td>D1-A4c</td>
<td>D4-A8c</td>
<td>8.18%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel (Ref 6)</td>
<td>D1-A1</td>
<td>D2-A4c</td>
<td>D4-A8c</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel (Ref 9)</td>
<td>D1-A2</td>
<td>D3-A4c</td>
<td>D4-A8c</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel (Ref 2)</td>
<td>D1-A4a</td>
<td>D4-A8c</td>
<td>3.72%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel (Ref 1)</td>
<td>D1-A4b</td>
<td>D4-A8c</td>
<td>6.81%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel (Ref 4)</td>
<td>D1-A1</td>
<td>D2-A4b</td>
<td>D4-A8c</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel (Ref 7)</td>
<td>D1-A2</td>
<td>D3-A4b</td>
<td>D4-A8c</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A5</td>
<td>D4^A8c</td>
<td>41.41%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A1</td>
<td>D2^A5</td>
<td>D4^A8c</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A2</td>
<td>D3^A5</td>
<td>D4^A8c</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A5</td>
<td>D4^A7</td>
<td>D6^A8c</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A5a</td>
<td>D4^A8c</td>
<td>3.75%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A8c</td>
<td>19.05%</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A1</td>
<td>D2^A8c</td>
<td>0.09%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A2</td>
<td>D3^A8c</td>
<td>0.15%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A4c</td>
<td>D4^A8</td>
<td>0.36%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A4b</td>
<td>D4^A8</td>
<td>0.10%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A5</td>
<td>D4^A8</td>
<td>0.71%</td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A8</td>
<td>0.55%</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Novel</td>
<td>D1^A8b</td>
<td>0.17%</td>
<td></td>
</tr>
</tbody>
</table>

*R* Reads used in these calculations came from HIV-1 sequence reads that had been amplified with the indicated primer pairs (F1.1/R1.3, F1.4/R1.4, or F4.1/R4.1). Nef 2 was used to normalize reads from primer pair 4.1 for comparison to ~2kb transcripts. Primer pair 4.1 was not used to amplify messages produced in HOS-CD4-CCR5, therefore transcripts identified with this primer pair could not be compared in HOS cells.

**Naming of transcripts is according to convention established in Purcell and Martin, 1993. Novel Tat^8c and Ref encoding transcripts were named similarly. New Nef transcripts were named by extension of the convention. Previously detected unnamed transcripts are left blank.

*** "I" indicates continuous exon. "*" indicates removed intron. Site labels are as in Table 5-3.

****Transcripts detected in less than 5 sequence reads in T cells.

transcripts use a set of little studied splice acceptors located ~800 base pairs downstream [9,13]. We determined the complete structure of 29 members of the 1kb class (Fig. 5-3 and Table 5-5). The most abundant messages in this class use the
novel acceptor A8c to define their terminal exon. For HIV\textsubscript{89.6}, the 1kb class of transcripts was nearly as abundant as the 2kb class (Table 5-3, compare splicing to A7 and A8-like acceptors), and this was supported by endpoint RT-PCR analysis (data not shown).

Acceptor A8c is not well conserved in HIV-1/SIVcpz (14%), although it is conserved in clade G viruses (>95%) and most HIV-2/SIVsmm genomes (86%) [45]. However, A8c is displaced from nearby well conserved (>90%) cryptic acceptors A8a and A8b by multiples of 3 base pairs (12 and 21 bp respectively), so splicing to any of these three acceptors would create similar ORFs. All HIVs and SIVs maintain at least one of these three acceptors, suggesting possible function [45]. We confirmed that the 1kb transcripts using A8a, A8b, and A8c were present in infected HOS and T cells by endpoint PCR and Sanger sequencing of cloned transcripts (Fig. 5-4A and data not shown).

The 1 kb transcript containing exons 1, 4 and 8c (1-4-8c, where exon 8c begins at A8c and extends to the poly-adenylation site) encodes the first exon of Tat followed by 25 novel amino acids (termed Tat\textsuperscript{\^8c}). Tat\textsuperscript{\^8c} showed activity when overexpressed in cells containing a Tat reporter construct (Fig. 5-4B, nucleotide and amino acid sequences in Table 5-6). Transcripts with exon structures 1-4a/b/c-8c, encode a novel fusion of Rev and Nef, hereafter referred to as Ref. Ref contains the amino-terminal 26 amino acids of Rev corresponding to the monomer’s first of two alpha helices, which are thought to be involved in folding, and oligomerization of
Figure 5-4 - Novel transcripts utilizing acceptor A8c. (A) HIV-189.6 cDNAs were amplified from HOS cells with primers keo056 and keo057 (black arrows). Major cDNAs were cloned for Sanger sequencing confirming PacBio® sequencing: HIV-1 genes, open arrows; start codons, circles; exons, thick bars; introns, dashed lines. Coding potentials of novel open reading frames (ORFs) are described at right. The first two start codons in messages 5 and 6 are not shared by known HIV-1 genes. Messages 1, 2, 4, and 5 were cloned into expression plasmids upstream of an internal ribosomal entry site and GFP. (B) Tat activity measured in Tzml-bl cells as Tat-dependent luciferase production after transient transfection with expression plasmids. GFP expression was equivalent for all constructs (microscopy, not shown). (C) Rev activity measured after co-transfection of HEK 293T cells with expression plasmids and a Rev-dependent Gag/Gag-Pol reporter plasmid. Intracellular Gag was assayed by FACS, gating on GFP positive cells. Rev and Ref were expressed alone or mixed (constant Rev, increasing Ref). Only the rev cDNA encoded Rev activity. (D) Infection efficiency (HIV-189.6 in HOS-CD4-CCR5 cells transfected with plasmids containing novel ORFs. Released Gag (p24) was measured at 72hpi (ELISA). Only Tat stimulated p24 production. (E) Expression of predicted Ref protein (12.5 kDa) from cloned cDNA in MG132 (proteosome inhibitor) treated cells, detected by Western blot using an antibody against the carboxy-terminus of Nef. Nef was expressed to control for production of partial Nef peptides or breakdown products from the Nef ORF.
### Table 5-6 - ORFs and amino acid sequences of Tat^8c and Ref.

<table>
<thead>
<tr>
<th>Transcript</th>
<th>Exons</th>
<th>cDNA ORF and Amino Acid sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tat^8c</td>
<td>1-8c</td>
<td>ATG GAG CCA GTA AAT CCT AGC CTA GAG CCC TGG AAG CAT CCA GGA AGT CAG CCT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M E P V N P S L E P W K H P G S Q P</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AAA ACT GCT TGT ACC AAT TGG TAT TGC AAA AAA TGT TGC TTT CAT TGC CAA GCT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>K T A C T N C Y C K K C C F H C Q A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TGT TTC ATA ACA AAA GGC TGA GGC TCC TAT GGC AGG AAG AAG CGG AGA CAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C F I T K G L G I S Y G R K K R R Q</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGA CGA AGA CCT CCT CAA GAC AGT CAG ACT CAT CAA GTT CTA TCA AAG CAA A/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R R R F P P Q D S Q T H Q V S L S K Q</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TTG GCA GAA CTA CAC ACC AGG GCC AGG AAT CAG ACA TCC ACT GAC CTT TGG ATG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>L A E L H T R A R N Q I S T D L W M</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GTG CTA CAA GCT AGT ACC AGT TGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>V L Q A S T S *</td>
</tr>
<tr>
<td>Ref</td>
<td>1-4c/a/b 8c</td>
<td>ATG GCA GGA AGA AGC GGA GAC AGC GAC GAA GAC CTC CTC AAG ACA GTC AGA CTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M A G R S G D S D E D L L K T V R L</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATC AAG TTC TCT TAT CAA AGC A/A AT TGG CAG AAC TAC ACA CCA GGA CCA GGA AGA ATC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>I K F L Y Q S N W Q N Y T P G P G I</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AGA TAT CCA CTG ACC TTT GGA TGG TGC TAC AAG CTA CCA GTT GAG CCA GAT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R Y P L T F G W C Y K L V P V E P D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GAA GGA AAC AAG AGA GAG GAC AAG AGG TGG CTA CAC CCT GCT AAC CAG CAT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>E G E N N R E D N S L L H P A N Q H</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GGA GTA GAA GAC TCC GAG AGA CAA GTG TTA GTG TGG AGG TTT GAC AGC GCG CTA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>G V E D S E R Q V L V W R P D S R L</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCA TTC CAT CAC GTG GCC CAA GAG CTG CAT CCG GAG TAG TGC ATG AAG AAC TGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A F H H V A R E L H P E Y F K N *</td>
</tr>
</tbody>
</table>

Rev [46]. These amino acids are fused to the carboxy-terminal 80 amino acids of Nef, required for Nef-dependent CD4 internalization and sorting to clathrin-coated pits [47]. We did not detect Rev activity on overexpression of the ref transcript, and Ref did not appear to interfere with the normal function of Rev or with HIV replication (Figs. 5-4C and D). Ref was detectable by Western blot using antibodies targeting the C terminus of Nef after inhibition of the proteosome, suggesting that the fusion is expressed but not stable (Fig. 5-4E). Thus Ref has the potential to encode a new epitope potentially relevant in immune detection of HIV. The transcripts with exon structures 1-5-8c and 1-8c encode at most a short peptide, and so are candidates for acting as regulatory RNAs.
Additional Novel Transcripts

Additional detected novel transcripts are noteworthy, though these occurred at low abundance and have not been further studied. The first major conserved splice acceptor in HIV-1 is A1 and, together with donor D2, it defines the boundaries of exon 1. In all previous examples of splicing from D2, it has been in the context of a fully spliced exon 2, and it has been shown that splicing at D2 helps drive splicing from D1 to A1 via exon definition [48]; however, our data showed for the first time that splicing from D2 to several downstream acceptors can occur without upstream splicing to A1. We could not distinguish whether these transcripts contained an intron upstream of D2 or whether the D2 splice junctions were the first splice events in these transcripts. The latter splicing pattern is particularly interesting as it would create spliced transcripts containing the \( \psi \) packaging signal (normally disrupted by splicing at D1), although it is likely repressed [49]. Packaging of such subgenomic RNAs would poison HIV replication.

Another transcript of interest contains the first coding exon of tat (exon 3) spliced to a novel small exon within env and in frame with env, with an additional splicing event to the final major HIV-1 exon (exon 7) again in frame with env. This fusion is reminiscent of the Tat-Rev-Env fusion, p28\textsuperscript{lev} (Tev), first described by Benko and colleagues; however, the 5’ boundary of the small central exon in Tev is an acceptor (A6a) found in HXB2 but not conserved [8,11]. The acceptor used in our novel proposed Tat-Env-Env message, dubbed A6b, lies downstream of A6a and is also not well conserved—it is found sporadically in most HIV-1 clades. Both fusions require splice donor D5, which is conserved in several clade B viruses, but not in other clades.
Finally, we detected a putative splice donor within exon 7, which we have termed D6. Due to our experimental design, we could only detect splice junctions involving this donor and the cluster of A8 acceptors. However, it is conceivable that if RNA Pol II read through the poly-adenylation site into adjacent host genomic DNA, this splice donor could be paired with host splice acceptors to create chimeric viral-host transcripts as has been observed with lentiviral gene-therapy vectors [50].

**Temporal Dynamics of Transcript Populations**

To assess longitudinal variation, we investigated HIV-1\textsubscript{189,6} transcript populations during the course of a single round of infection in HOS-CD4-CCR5 cells. A sensitive method for comparison among conditions involves quantifying utilization of six mutually exclusive splice acceptors A3, A4c, A4a, A4b, A5, and a novel acceptor just downstream of A5 termed A5a. Splicing at these acceptors determines the relative levels of messages encoding Tat and Env/Vpu in the partially spliced class, and messages encoding Tat, Rev, and Nef in the completely spliced class.

We observed longitudinal changes in the levels of these messages in HOS cells over 12-48 hours that were statistically significant (p<10\textsuperscript{-10}; generalized linear model described in methods, analysis by Scott Sherrill-Mix). This pattern was especially evident in junctions involving donor 1 spliced to each of these acceptors (Fig. 5-5A). Most dramatically, transcripts with splicing junctions between D1 and A3 (tat messages) increased with time (p<10\textsuperscript{-10}), while D1\textsuperscript{A}A4b junctions (used in
env/vpu or rev messages) were used reciprocally less (p<10⁻¹⁰). Such kinetic
changes affecting specific transcripts both with and without the Rev-response
element cannot be explained by the accumulation of Rev, and they may reflect
differential transcript stability or HIV-induced alterations to the host splicing
machinery. Temporal changes in HOS cells were confirmed using endpoint RT-PCR
and analysis after electrophoresis on ethidium-stained gels (Fig. 5-5B).

Cell-Type Specific Splicing Patterns

We also compared splicing between T cells and HOS cells and found
significant cell type differences (p<10⁻¹⁰). For example, while transcripts with
D1^A5 junctions were dominant in both cell types, messages using the D1^A4c
splice junction (encoding Env/Vpu or Rev) made up the bulk of the remaining
transcripts in T cells but were a minor species in HOS-CD4-CCR5 cells. Likewise, Tat
messages (using A3), which were quite abundant in HOS cells at all time points,
contributed relatively little to populations of transcripts in primary T cells
harvested at 48 hpi (compare Fig. 5-5A). We also used end point PCR and analysis
on ethidium bromide-stained gels to confirm that the relative ratios of transcripts
containing junctions to A3, A4a, A4b, and A4c were different in HOS and T cells (Fig.
5-5B).
Human Variation in HIV-1 Splicing

Quantitative comparisons also revealed modest differences in splicing between primary CD4+ T cells isolated from different human donors that were statistically significant (p<10^{-10}) under a generalized linear model (Fig. 5-5A). The magnitudes of predicted differences were small, all less than 33% and most lower than 10%, but quite reproducible.

Figure 5-5 - Temporal, cell-type, and donor variability in accumulation of HIV-1 messages. (A) To highlight changes in ratios of HIV-1 transcripts over time and between HOS-CD4-CCR5 cells and primary T cells, we used PacBio® sequence read counts to calculate proportions of transcripts with splicing from D1, to each of the mutually exclusive splice acceptors: A3 (required to make Tat), A4c, A4a, A4c, (Env/Vpu and Rev), and A5 (Env/Vpu and Nef). Sequences derived from templates amplified with primers F1.2 and R1.2 (Table 5-1). Heat map shows average data for HOS and T cell samples in columns, with the color tiles indicating proportion of D1 splicing to each mutually exclusive acceptor (rows), according to the color scale shown. (B) Verification of temporal and cell type changes shown in (A) by bulk RT-PCR amplification of HIV-1_{89,6} transcripts from HOS cells and one human donor’s primary T cells (subject 3). Resolved by gel electrophoresis and stained with ethidium bromide.
Discussion

Single molecule enrichment and long read single molecule sequencing have made possible the most complete study to date of the composition of HIV-1 message populations, revealing several new layers of regulation. Studies of the low-passage HIV$_{89.6}$ isolate in a relevant cell type showed numerous differences from studies of lab adapted HIV strains in transformed cell lines, highlighting the importance of studying the most relevant models. These data also illustrate the limitations of gel-based assays for studying HIV-1 message populations. Multiple different combinations of HIV-1 exons yield mRNAs of similar sizes that are easily confused in typical assays using gel electrophoresis. Thus in many settings the more detailed information provided by single molecule amplification and single molecule DNA sequencing is more useful.

Using these methods, we have detected significant variations between HIV message populations generated in T-cells from different human donors. The differences were modest compared to those observed between cell types or time points, perhaps not surprisingly since any human polymorphisms strongly affecting mRNA processing might interfere with normal gene expression. However, because tight calibration of message levels is important to HIV-1, the observed differences in message ratios might affect HIV-1 acquisition or disease progression. Given the temporal changes observed in HOS cells, the variation in observed transcripts could also be affected by different kinetics of infection in T cells from the different donors. In either case, these data suggest that human polymorphisms may exist that affect
HIV-1 message populations in infected individuals, providing a new candidate mechanism connecting human genetic variation with measures of HIV disease.

Sequences from the 89.6 viral strain revealed a class of small (~1kb) completely spliced transcripts, most contributed by splicing to a new poorly conserved acceptor A8c. These encoded two new proteins, one of which had Tat activity, and we showed that another, a Rev-Nef fusion termed Ref, could be detected in cells. HIV-1_{89.6} is a particularly cytotoxic virus isolated from the CSF of a patient, and it forms unusually large syncitia in macrophages [31]. The abundance of 1kb transcripts produced by this virus provides a possible explanation for its unique properties. In addition to the novel acceptor A8c, we have also identified 3 putative novel splice donors and 11 putative novel acceptors, which must be further studied for possible functions.

The wealth of new messages found here in HIV-1_{89.6} and in other HIV-1 isolates suggests there may be ongoing evolution of novel splice sites and new ORFs. Because splice acceptors in HIV-1 are weak [14], mutations creating sequences that even slightly resemble the 3’ splice site consensus may be occasionally recruited as novel acceptors, creating new transcripts. In fact, new splice signals may evolve with relative ease – it has been estimated that reasonable matches to the consensus for splice donors, acceptors and branch-point sites occur within random sequence every 290, 490, and 24 base pairs respectively [51].

We and others have observed appearance of novel exons within the major HIV-1 introns [8,10,11]. Such long stretches of RNA relatively devoid of competing
splice sites may be particularly poised to evolve new signals. On the other hand, most of the putative novel splice acceptors we observed clustered near previously identified acceptors in HIV-1, suggesting that conserved cis-acting splicing signals may recruit factors that act promiscuously on new nearby sequences. Interestingly, clusters of novel sites that would create redundant mRNAs were found near D1 and A5, which are both used in large proportions of HIV-1 transcripts. Particularly strong splice enhancers may drive the use of the novel sites. Clusters of splice sites might be evolutionarily beneficial for the virus, providing redundancies in these locations to protect vital messages [52]. Along these lines, Abbink and Berkhout have reported that, in viruses where the use of the major donor in the upstream leader, D1, is compromised, mutation of a GC within the dimerization initiation sequence (DIS) stem to GT can rescue splicing and replication of HIV-1 [52]. This GC is just 7 nucleotides upstream of the GC that defines D1b, also found within the DIS. We speculate that both upstream GCs can act as “backup” donors for D1. In another particularly interesting example of the evolutionary flexibility conferred by splice site redundancies, Dr. James Hoxie has observed rescue of an SIVmac bearing incapacitating mutations in env by second-site mutations abolishing the major splice acceptor equivalent to A7. These revertant mutations would themselves cripple the virus if it were not for the presence of a redundant splice site conserved in SIVmac just upstream of A7, which substitutes to produce normally balanced mRNA populations in the revertants (unpublished data).
Frequent evolution of new splice sites may also allow viruses to test out new combinations of exons with the potential to create novel functional RNAs or proteins, like those reported here. However, such novelty must compete with immune constraints – unstable novel polypeptides like Ref can be targeted to the proteosome and presented on MHC molecules as new epitopes for immune recognition. Though Ref consists of portions of Rev and Nef, the junction of the two protein components may provide such an epitope. The short peptides predicted to be encoded by other ~1kb transcripts might similarly provide new epitopes. Alternatively, these transcripts may function as RNAs. Supporting this idea, the RNA consisting of exons 1, 5, and 8c is predicted to be highly structured (not shown).

HIV has likely evolved to produce calibrated message populations in T cells, which seem to be altered with relative ease, as in infection in HOS cells, suggesting that therapeutic disruption of correct splicing may be feasible. A few studies have begun to explore small molecule therapy to disrupt HIV-1 splicing [15,19]. Several factors could be responsible for the differences we observed between HOS and T cells, including hnrNP A/B and H, SC35, SF2/ASF and SRp40 (discussed in Chapter 1) [53,54]. Inhibition of SF2/ASF has already been shown to abrogate HIV-1 replication in vitro [15]. Thus the lability seen here for function of these factors suggests they may be attractive antiretroviral targets.
Acknowledgements

We would like to thank the University of Pennsylvania Center for AIDS Research (CFAR) for preparation of viral stocks and isolation of primary CD4+ T cells; James A. Hoxie, Ronald G. Collman, Jianxin You, Robert W. Doms, Paul Bates, David Rekosh, and members of the Bushman laboratory for reagents, helpful discussion and technical expertise. This work was funded by the HINT consortium (P01 AI090935). K.E.O. was supported by T32 AI 7324-19.

References


CHAPTER 6 – DYNAMIC HIV-1 MESSAGE POPULATIONS AND CELL CYCLE ARREST

Abstract

Populations of HIV-1 mRNAs are tightly calibrated to produce the correct relative levels of the viral proteins. We have observed in HOS cells that mRNAs encoding Tat increase with time, concomitant with a decrease in other mRNAs. Nocodazole treatment causes a similar shift in mRNA populations in HOS and CD4+ T cells, therefore we speculate that G2/M phase cell cycle arrest or apoptosis, potentially induced by HIV-1 Vpr expression, trigger changes in activity or relative levels of cellular splicing factors, leading to the observed changes in mRNA populations. However we did not see reproducible differences in mRNA populations between wild type virus and HIV-1 carrying nonsense mutations in Vpr. Dynamic differences in viral RNA populations between cell types suggest that tissue specific programs of splicing and RNA degradation exist and may have driven evolution of HIV-1 mRNA regulatory features.

Introduction

Gene expression is controlled at several steps. While much focus is placed on transcription and translation, the intermediate processes of mRNA splicing and regulated RNA degradation play equally important roles. Cellular mRNAs are post-transcriptionally modulated to drive cellular processes as fundamental as cell cycle
progression: regulated dephosphorylation of the splicing factor, SRp38, alters cellular splicing programs to facilitate repression of interphase gene expression during mitosis [1], and cell cycle progression is promoted by degradation of mRNAs encoding histones and cyclins [2]. Cell-cycle arrest often triggers apoptotic programs, and this switch has also been linked to changes in splicing resulting in expression of the pro- rather than anti-apoptotic forms of Mcl1 and Bclx [3].

Global cellular changes in splicing [4] and mRNA degradation [5] are also observed upon stimulating resting T cells. For example, increased levels of the splicing regulator hnRNP-LL upon T cell activation are associated with skipping of several exons in messages of CD45, yielding a less active isoform of the protein tyrosine phosphatase, which helps to maintain T cell homeostasis [6]. Engagement of CD28 on the T cell surface leads to stabilization of mRNAs for interleukin-2, interferon-γ, granulocyte-macrophage colony stimulating factor, and tumor necrosis factor-α [7].

These two RNA-level regulatory mechanisms also influence expression of viral genes in infected cells. HIV-1 preferably infects and replicates in activated CD4+ T cells, and HIV-1 infection leads to global immune stimulation [8]. Once a T cell is infected, Vpr induces cell cycle arrest at the G2 phase [9,10], likely by targeting a factor that promotes cell cycle progression for degradation by the cullin4A-DDB1-DCAF1 E3-ubiquitin ligase complex [11-13]. This eventually induces cell death [14,15]. In the midst of what must be a dynamically changing environment of cellular splicing and mRNA degradation programs, HIV must
express its proteins in appropriate proportions so as to efficiently assemble new infectious virions.

As described in Chapter 1 of this dissertation, the HIV-1 pre-mRNA interacts with several host splicing regulators through a variety of exonic and intronic cis-acting splicing enhancers and suppressors to calibrate the populations of mRNAs produced [16]. For example, levels of tat messages are kept low by several splicing repressors whereas env/vpu, rev, and nef messages are promoted by the presence of enhancers [16]. In macrophages, HIV-1 infection may modulate cellular splicing factors, resulting in a decrease in Tat production that enables formation of quiescent reservoirs of HIV-1 in these cells [17].

The relative stability of the HIV-1 mRNAs also modulates viral gene expression. Unspliced and partially spliced mRNAs remain in the nucleus and are degraded until Rev, expressed from a completely spliced transcript, accumulates to export them from the nucleus, thus mediating the temporal progression of HIV-1 gene expression [18,19]. Furthermore, it has been shown that for at least 3 completely spliced messages, the HIV-1 mRNAs have differing half-lives (Tat<Rev<Nef) [17].

We have observed that the relative proportion of HIV-1 mRNAs contributed by tat messages increases with time post infection in a line of human osteosarcoma cells (HOS) expressing HIV-1 receptors CD4 and CCR5 (see Chapter 5). Here we present data in HOS cells and primary human CD4+ T cells showing that this shift can be mimicked with Nocodazole treatment, leading us to hypothesize that it may
be tied to cellular changes of G2/M phase cell cycle arrest or apoptosis. However, the increase in tat messages is not dependent on Vpr expression, and may in fact be repressed by it, indicating that other mechanisms may be at play. Cell-type specific differences were also observed. These data suggest that HIV-1 is particularly adapted to produce calibrated message populations under splicing and mRNA degradation programs active in stimulated CD4+ T cells, and further, that the virus might modify cellular programs in T cells to its own benefit, much as it does in macrophages [17].

Methods

Cell culture and Infections

Low passage (<20) Human osteosarcoma cells expressing CD4 and CCR5 (HOS-CD4-CCR5, AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH, from Dr. Nathaniel Landau [20,21]) were grown and passaged in DMEM with GlutaMAX (Invitrogen) supplemented with 10% FBS (Sigma-Aldrich), and 1 μg/mL Puromycin (Calbiochem). Primary CD4+ T cells were isolated by the University of Pennsylvania Center for AIDS research Immunology core from apheresis product using the RosetteSep Human CD4+ T Cell Enrichment Cocktail (StemCell Technologies). T Cells were stimulated for 3 days prior to infection at 0.5 x10^6 cells per milliliter in R10 media (RPMI 1640 with GlutaMAX (Invitrogen) supplemented with 10% FBS (Sigma-Aldrich) with 100 units/mL recombinant IL2 (Novartis) + 5ug/mL PHAL (Sigma-Aldrich).
All viruses were generated by the University of Pennsylvania Center for Aids Research by transfection of 293T cells with plasmid clone, and passage in SupT1 cells. Infections were performed with virus equivalent to 100-200 ng p24 per million cells. For HOS cell infections, virus was incubated with cells for 4 hours at 37°C and then replaced with fresh media following a PBS wash. For CD4 T cell infections, cells were spinoculated 2 hours at 37°C at 1200xg and then incubated for an additional 2 hours at 37°C prior to addition of new media. For single round infection, Nelfinavir or Enfuvirtide was added at 6-12 hpi (T-20, Fusion Inhibitor, AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH,).

RNA and Reverse Transcription

Total cellular RNA was purified using the Illustra RNA kit (GE Life Sciences, Fairfield, CT) from 0.5-5x10^6 cells per infection depending on the study. Transcript populations were amplified from cellular RNA using the Onestep RT-PCR kit (Qiagen) as follows: 35 cycles of 30s at 94°C, 30s at 56°C, and 45s at 72°C followed by 10 min at 72°C. Primers F1.2 and R1.2 were used for amplification of partially and fully spliced transcripts together. Primers F1.3 and R1.3 were used to amplify selectively incompletely spliced transcripts. Primers F1.4 and R1.4 were used to investigate ratios of completely spliced transcripts. Oligonucleotides used in this study are listed in Table 6-1 and depicted with respect to the HIV-1_89.6 genome in Chapter 5 of this dissertation (Fig. 5-1). Products were resolved on agarose gels
(Invitrogen or Nusieve 3:1, Lonza for better resolution) and stained with ethidium- bromide (Sigma) for visualization.

Table 6-1 – DNA Oligonucleotides used in this study.

<table>
<thead>
<tr>
<th>ID</th>
<th>Sequence (5'---&gt;3')*</th>
<th>Description</th>
<th>HIV-189.6 Coordinates</th>
<th>Split?</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1.2</td>
<td>CGTCTGGTTGTGACTCTGG</td>
<td>Forward primer, to compare use of splice acceptors A1-A5 in all transcripts.</td>
<td>564:583</td>
<td></td>
</tr>
<tr>
<td>R1.2</td>
<td>AGGTCTTCGTCGTCTTC</td>
<td>Reverse primer, to compare use of splice acceptors A1-A5 in all transcripts.</td>
<td>5984:6002</td>
<td></td>
</tr>
<tr>
<td>F1.3</td>
<td>CAAGTAGGTGTGAGCGTC</td>
<td>Forward primer, to compare use of splice acceptors A1-A5 in partially spliced transcripts</td>
<td>549:567</td>
<td></td>
</tr>
<tr>
<td>R1.3</td>
<td>GTGGTGGTCTTCTTC</td>
<td>Reverse primer, to compare use of splice acceptors A1-A5 in partially spliced transcripts</td>
<td>6353:6371</td>
<td></td>
</tr>
<tr>
<td>F1.4</td>
<td>GATCTCTCGAGGCGAGGCT</td>
<td>Forward primer, to compare use of splice acceptors A1-A5 in completely spliced transcripts</td>
<td>679:697</td>
<td></td>
</tr>
<tr>
<td>R1.4</td>
<td>GGGAGGAGGGTGGTTT</td>
<td>Reverse primer, to compare use of splice acceptors A1-A5 in completely spliced transcripts</td>
<td>6038:6043; 8368:8378</td>
<td>s</td>
</tr>
</tbody>
</table>

Coordinates refer to HIV-189.6 reference sequence. Primers that span an exon-exon junction are designated as split ("s") and the coordinates reflect the removed intron.

Cell Cycle inhibitors and Apoptosis Assays

For studies using cell cycle inhibitors, infections were performed as above in HOS-CD4-CCR5 cells in triplicate or primary human CD4⁺ T cells in duplicate and allowed to proceed 21 hr (HOS) or 48 hr (CD4⁺ T) prior to treatment. For each drug, uninfected cells in six wells (HOS) or 3 wells (CD4⁺ T) of a 96 well plate were treated to assess induction of apoptosis. Cultures were then incubated with the following inhibitors in D10 media for 18hr prior to harvest of infected cells or apoptosis assays: Adriamycin (Doxorubicin hydrochloride, Sigma) at 2 μM, Nocodazole
(Sigma) at 200 μM, Thapsigargin (Sigma) at 1 μM and Staurosporine (Sigma) at 1 μM. Each inhibitor was stored in DMSO and a carrier control was performed with a volume of DMSO equivalent to the volume of the least diluted drug in HOS cells or matching the volume of each drug in CD4+ T cells. Induction of apoptosis was assayed using the Anaspec SensoLyte® Homogeneous Rh110 Caspase-3/7 Assay Kit according to manufacturer’s directions.

Results

Temporal dynamics of transcript populations

In order to track dynamic changes of HIV-1 mRNA populations over time, we amplified completely and incompletely spliced transcripts from HOS-CD4-CCR5 cells infected with the early passage isolate HIV-189.6. The infection was restricted to a single round by the addition of anti-retroviral drugs. As shown in Figure 6-1, we saw an increase in abundance of tat transcripts with time as described in Chapter 5 (see Fig. 5-5). The change was observed most strongly in the incompletely spliced class of transcripts (Fig. 6-1A), but it was weakly observed in the completely spliced class as well (Fig. 6-1B), suggesting that splicing at acceptor A3, used to make tat transcripts in both classes, might increase with time. The kinetics of this change differed, and they were notably slower when the fusion inhibitor Enfuvirtide was used in place of Nelfinavir to prevent spreading infection (Fig. 6-1, discussed below). A variable decrease in smaller transcripts (encoding Env/Vpu or Rev and Nef depending on downstream splicing) accompanied the increase in tat mRNAs in
several experiments. This work was done with Steven Siegel and Edward Kreider, University of Pennsylvania graduate students.

**Figure 6-1 - Increased relative abundance of HIV-1<sub>189,6</sub> mRNAs encoding Tat with time post infection in HOS-CD4-CCR5 cells.** Infection was restricted to a single round with (A) Nelfinavir (protease inhibitor) and (B) Enfuvirtide (fusion inhibitor). RT-PCR was performed on equal amounts of total RNA from each sample. Increase in tat transcripts is more pronounced within the partially spliced (amplified with primers F1.3 and R1.3, upper panel) than in the completely spliced class of transcripts (amplified with primers F1.4 and R1.4, lower panels), and accelerated in the Nelfinavir-treated infection compared to the Enfuvirtide-treated infection. This work was done with Steven Siegel (A) and Edward Kreider (B).

**Cell cycle inhibitors induce changes to HIV-1 transcript populations**

While the kinetics of the increase in tat transcripts varied, the shift appeared to correspond temporally with a decrease in cell viability in the culture. Changes in splicing of transcripts encoding Mcl1 and Bclx resulting in a shift from expression of the anti- to pro-apoptotic protein forms are stimulated by cell cycle arrest in G2/M phase and contribute to induction of apoptosis [3]. We hypothesized that changes in cellular splicing factors during cell cycle arrest may similarly affect splicing of HIV-1 RNAs. Our hypothesis was supported by the noted more rapid increase in tat messages in the presence of Nelfinavir, which induces cell cycle arrest and apoptosis in several transformed cell types [22-24] (Fig. 6-1A and B and Fig. 5-5).
In order to test whether cell cycle arrest could induce changes in HIV-1 mRNA populations, we investigated splicing of HIV-1\textsuperscript{89,6} transcripts under treatment with a panel of inhibitors (Staurosporine, Nocodazole, Thapsigargin, and Adriamycin) that arrest the cell cycle by different mechanisms and at different phases, and induce apoptosis (Table 6-2) [25-34]. Spreading infection was inhibited by Enfuvirtide, and transcript ratios were examined at 39 hpi, a time-point when similarly infected cells had not yet exhibited the increase in \textit{tat} transcripts. Activity of Caspase 3 and Caspase 7 was measured as an indicator of induction of apoptosis (Fig. 6-2A). Cellular responses varied with Nocodazole and Staurosporine inducing the most cleavage of caspases 3 and 7 and the most cell death in HOS-CD4-CCR5 cultures as observed by microscopy (not shown). As shown in Figure 6-2B, the ratio of \textit{tat:nef} transcripts was increased in cells treated with Nocodazole, when compared to the relative levels of the transcripts in carrier-treated cells. This shift towards increased \textit{tat} mRNAs was similar to that observed late in HOS cell infections and was not observed with the other cell cycle inhibitors. Similar results were observed in primary human CD4\textsuperscript{+} T cells (Figs. 6-2C and D).

\textbf{Table 6-2 – Cell cycle inhibitors used in this study.}

<table>
<thead>
<tr>
<th>Inhibitor</th>
<th>Mechanism</th>
<th>Phase of cell cycle arrest</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staurosporine</td>
<td>general kinase inhibitor</td>
<td>Evidence for G1 and G2</td>
<td>Crissman \textit{et al.} [29], Abe \textit{et al.} [25], Bruno \textit{et al.} [28]</td>
</tr>
<tr>
<td>Nocodazole</td>
<td>Inhibits microtubule polymerization</td>
<td>G2/M</td>
<td>Zieve \textit{et al.} [34], Nusse and Enger [31]</td>
</tr>
<tr>
<td>Thapsigargin</td>
<td>Perturbs cellular calcium and induces ER stress</td>
<td>Evidence for G1 and G2</td>
<td>Simon and Moran [33], Bourougaa \textit{et al.} [27]</td>
</tr>
<tr>
<td>Adriamycin</td>
<td>Intercalates in DNA and inhibits Topoisomerase II</td>
<td>S (predominantly)</td>
<td>Kim and Kim [30], Barranco \textit{et al.} [26], Pommier \textit{et al.} [32]</td>
</tr>
</tbody>
</table>
Figure 6-2 - Treatment of infected cells with Nocodazole results in increased relative levels of *tat* mRNAs. (A) Induction of apoptosis by 18h treatment with cell cycle inhibitors measured as activity of caspases 3 and 7 on a substrate which is cleaved to produce a fluorescent product. All measurements were normalized to those in cells treated with DMSO carrier. Inhibitors Staurosporine (Stauro), Nocodazole (Nocod), Thapsigargin (Thaps), Adriamycin (Adri) were added to (B) HOS-CD4-CCR5 and primary (C) CD4+ T cells, 21 or 48 hours post HIV-189.6 infection respectively. Equal amounts of total cellular RNA were subject to RT-PCR. Primers F1.2 and R1.2 were used to amplify all transcripts without discriminating between completely and partially spliced mRNA classes (Table 6-1 and Fig. 5-1). Partially spliced mRNAs were amplified by primers F1.3 and R1.3.

Ratio of *tat* mRNAs to other viral mRNAs is not dependent on Vpr in HOS and T cells

Nocodazole induces cell cycle arrest at G2/M phase [31,34]. Because the HIV-1 accessory protein Vpr also induces cell cycle arrest at G2 phase [9,10], we
speculated that the increase in the relative abundance of *tat* mRNAs might be secondary to Vpr-dependent cell cycle arrest and cell death. To investigate this possibility, we examined changes in mRNA transcript populations over time, expressed from wild-type virus and HIV-1<sub>89,6</sub> containing nonsense mutations in Vpr. Two mutant viruses were used: HIV-1<sub>89.6-dVpr5125/5134</sub> (nonsense mutations in the 8<sup>th</sup> and 11<sup>th</sup> codons of Vpr) and HIV-1<sub>89.6-dVpr5110</sub> (nonsense mutation in the 3<sup>rd</sup> codon of Vpr) [35].

Surprisingly, in HOS-CD4-CCR5 cells transfected with plasmids carrying HIV-1<sub>89.6</sub> and HIV-1<sub>89.6-dVpr5125/5134</sub> genomes, the increase in relative abundance of *tat* transcripts with time was more pronounced in the absence of Vpr (Figure 6-3A). This experiment was done by Edward Kreider, a graduate student at the University of Pennsylvania. This result contrasts the lack of increase in *tat* transcripts expected if Vpr-induced arrest were responsible for the change, and suggests that Vpr-induced changes in splicing or mRNA degradation may oppose a relative accumulation of *tat* messages with time.

Notably, we observed little effect of Vpr on viral mRNA populations in primary human CD4<sup>+</sup> T cells. In replicate experiments, there was no apparent difference between mRNA populations produced by the viruses over time (Fig. 6-3B). Vpr mutations were confirmed to be present by Sanger sequencing, and more rapid cell division was observed in cultures infected with the Vpr mutants (not shown). Replicate experiments were done by Rebecca Custers-Allen, a member of the Bushman lab. Our observations support the idea that cell-type specific
differences modulate splicing and/or degradation of HIV-1 mRNAs, consistent the idea that the role of Vpr may vary by cell type [11,35].

**Figure 6-3 - Role of Vpr in calibrating HIV-1 mRNA populations in HOS and CD4+ T cells.**

(A) HOS-CD4-CCR5 cells were transfected with plasmids carrying the wildtype (WT) or HIV-189.6-dVpr5125/5134 genome (ΔVpr(5125/5134)). Primers F1.2 and R1.2 were used to amplify transcripts without discriminating between completely and partially spliced transcript classes (see Table 6-1 and Fig. 5-1). Shift to increased tat mRNAs is more pronounced in the absence of Vpr. (B) Primary human CD4+ T cells were infected with wild type HIV-189.6 or virus carrying nonsense mutation(s) in Vpr, HIV-189.6-dVpr5125/5134 and HIV-189.6-dVpr5110 (ΔVpr(5110)). HIV-1 mRNAs were amplified as in (A). Although total HIV-1 mRNA recovered varied, proportions of the different transcripts remained fairly constant and did not differ between viruses. Hours post infection, hpi.
Discussion

We have shown that Nocodazole treatment induces increased production of Tat-encoding mRNAs in HIV-1 infected cells, mimicking effects seen late in infection of a human osteosarcoma cell line. HIV-1 Vpr, which like Nocodazole, arrests cells in G2 phase leading to cell death, did not appear to be required for increased tat message accumulation, and in fact Vpr may suppress increased tat expression in HOS cells.

It is possible that the effects of Nocodazole were due to its direct role in preventing microtubule polymerization rather than the induction of cell cycle arrest, suggesting a role for the cytoskeleton in controlling mRNA populations. The extensive interactions of HIV-1 with the cytoskeleton [36] could alter trafficking of mRNA-processing machinery in the cell. However, the similarity between the changes seen in HIV-1 message populations in HOS cells in the absence of Vpr and under Nocodazole treatment strengthens the tie of the effects to cell cycle progression or cell death. Notably, it has been suggested that the Env subunit, gp120, re-structures the cytoskeleton, and as a result, sensitizes the cell to apoptosis [36]. Thus, cytoskeletal changes and the cell cycle are intertwined, especially in the context of HIV-1 infection, confounding our interpretation.

An increase in tat messages may be caused by decreased relative degradation of those messages, increased production by means of splicing at the tat-specific splice acceptor, A3, or a combination of the two. Addressing the first possibility, tat messages may be more stable and accumulate with time in HOS cells and in
Nocodazole-treated cells. However, the stabilities of several HIV-1 mRNAs have been measured in macrophages, and the half-life of tat messages is unchanged with time and shorter than that of rev and nef messages [17]. Therefore this model requires there to be cell-type specific differences in relative stability. While this might be true, the cell-type specific effects of Vpr further complicate this model.

We favor a model in which differences in splicing account for the changes seen. In isolation, the splice acceptor A3 is strong relative to other acceptors in HIV-1 [37]. However several cis-acting elements in the HIV-1 pre-mRNA repress splicing at A3 and enhance splicing at downstream sites such that mRNAs encoding Env/Vpu, Nef and Rev are more abundant in infected cells (see Chapter 5 and [16,37,38]). This suppressive arrangement may poise HIV-1 for rapid induction of Tat production. The major exonic splicing silencer, ESS2, located 70 nucleotides downstream of A3 binds hnRNP A/B, and further silencing is derived from hnRNP H interactions with ESS2p (see Chapter 1, Fig. 1-6). An exonic splicing enhancer that overlaps with ESS2 is bound by SC35 to promote splicing, or relieve inhibition by hnRNP A/B [16]. Any of these factors might be modulated with cellular changes to explain the data - perhaps in HOS cells one or more of the suppressive interactions is released. Vpr has previously been reported to decrease cellular splicing [39,40], and this is consistent with its apparent suppression of increased splicing at A3 in HOS cells.

Why might it be beneficial for HIV-1 to regulate Tat expression in an inducible fashion? Excess tat may induce overly-abundant production of viral
proteins, stressing the cell and/or activating immune responses, and it may prevent the establishment of latency in appropriate cells [41-43]. In addition, Tat is known to change cellular expression patterns, independently causing cytotoxicity [44-46]. For these reasons it might be advantageous to keep tat messages at low levels. However there may be situations in which it is advantageous to release the repression of Tat. For example, as an infected cell approaches death, reducing cytotoxicity may become less important and increased HIV-1 gene expression may enable additional virion production prior to apoptosis.

It remains unclear how these effects are mediated. It has been proposed that Vpr-induced arrest in G2 phase enables increased expression of HIV-1 genes [47], and it is possible that regulation of splicing factors contributes. However, other data suggests that arrest may be a byproduct of a more important role in combating innate immunity: Vpr interactions with the cullin4A-DDB1-DCAF1 E3-ubiquitin ligase complex likely lead to degradation of an innate immune factor in macrophages, and proteins regulating cell cycle progression and splicing may be lost as collateral [11]. Therefore, it will be of interest to investigate the levels and activities of splicing factors acting at the tat splice acceptor in the presence and absence of Vpr and/or Nocodazole. We note that transformed HOS cells may have important differences in cell cycle regulation, which might have contributed to cellular responses observed in these cells.

The data presented here reveal a complex inter-dependence of cellular programs affecting splicing and mRNA stability. We have used HIV-1 as a sensitive
system with which to probe these processes, which likely also modulate cellular mRNA populations and expression in important ways.

Acknowledgements

The work presented in this chapter involved significant contributions from Steven Siegel, Edward Kreider, and Rebecca Custers-Allen. We'd also like to thank members of the Bushman lab for helpful discussion.

References

CHAPTER 7 – CONCLUSIONS AND FUTURE DIRECTIONS

In this dissertation, I presented studies of the determinants of HIV-1 integration and subsequent mRNA processing in human cells. For both of these steps of retroviral replication, the patterns produced – in proviral distributions and the proportions of different HIV-1 mRNAs respectively – revealed new roles for cellular factors. Fully clarifying the mechanisms by which such cellular factors influence HIV-1 replication remains a goal for future studies.

Studies of HIV-1 Integration

HIV-1 preferably inserts its genome into active transcription units and in gene dense regions of the host cell chromatin. In chapters 2 and 3, I showed that components of the nuclear pore and specific HIV-1 capsid interactions with Ran-Binding Protein 2 (RanBP2) and Cyclophilin-A (CypA) are responsible in part for directing the pre-integration complex (PIC) to chromatin rich in genes and associated features. However, the exact mechanisms by which the nuclear pore directs the PIC to gene dense regions, and how these mechanisms interface with other known determinants of integration site selection – for example tethering of IN to chromatin by PSIP1/LEDGF/p75 (LEDGF) – remain unclear. Such mechanisms, once explained, might help fill in the gaps exposed in Chapter 4 that remain in our knowledge of what makes certain sites and regions of the human genome especially favorable for normal HIV-1 integration.
It could be argued that disruption of the nuclear pore by the knockdowns presented in Chapter 2 simply prevents access of the PIC to the nucleus except during mitosis, and the more random distribution of integration sites in these cells reflects the vastly different organization and condensation of chromosomes in this phase. However, this would predict that, in cells arrested in interphase, integration events would show stronger targeting biases, undiluted by more random events occurring during mitosis. Though this was indeed observed in arrested IMR-90 cells [1], quite the opposite was observed in post-mitotic macrophages, where integration was slightly more random than in dividing cell lines [2]. It might be of interest to investigate the effects of depleting RanBP2 and TNPO3 and of HIV-1 capsid mutations in these cells. Though more difficult, reciprocal studies of integration events occurring in the absence of an intact nuclear envelope would also be informative. The SV-40 viral protein 4 (VP4) [3] and overexpressed myotonic dystrophy protein kinase (DMPK) [4] both destabilize the nuclear envelope, and may be useful in such studies. In addition, some human cancer cells have intermittent defects in nuclear envelope integrity during interphase [5].

Two models have been proposed to explain the role of the nuclear pore in integration. First, because there is evidence that some active genes relocate to the nuclear pore [6,7], it may be that passage through the pore delivers the PIC into a region of the nucleus enriched for gene dense chromatin. Consistent with this model, depletion of Transportin-3 and RanBP2 caused integration to decrease across broad gene-dense regions of chromatin losing association with all correlated
features, rather than alter preferences for particular features, as would be expected for depletion of a tethering factor, like LEDGF [8,9]. Also consistent, depletion of several nuclear factors disrupted integration in gene-dense regions, suggesting that nuclear organization is important to HIV-1 integration. Moreover, microscopy studies have demonstrated that PICs colocalize with open chromatin in the nuclear periphery, though it is uncertain where their genomes ultimately integrate [10].

However, apart from locations immediately adjacent to pores, the nuclear periphery is generally silencing for chromatin [7]. In Chapter 4, I showed that particularly attractive loci for integration were especially abundant on several small chromosomes, which are known to reside more frequently in internal regions of the nucleus [7]. Finally, in Chapter 2, I demonstrated that the HIV-1 CA mutant N74D, which traverses the nuclear pore in a RanBP2- and Transportin-3-independent manner [11], integrated in less gene dense chromatin than wild type virus, suggesting that interactions with particular components of the pore might be important. Therefore, a second model holds that the PIC sequentially engages specific host factors on a “railway track” to navigate directly to favored regions of chromatin. For example, RanBP2 might displace CypA from CA and pass the PIC to Transportin-3, with which, in turn, the PIC could traffic to regions enriched in splicing factors [12], and these might finally deliver the PIC to active genes. Supporting this model, unintegrated HIV-1 DNA was found to co-localize with SC-35, a serine and arginine rich splicing factor imported by Transportin-3 that clusters in central regions of the nucleus [13].
To examine these two models, it would be of interest to study the localization of proviral DNA within the nucleus with respect to the nuclear pore, for example, by fluorescence in-situ hybridization. ChIP studies to identify chromatin closely associated with the pore might also reveal correlations with integration distributions. In yeast, specific genes have been experimentally tethered to the nuclear pore with a resulting increase in transcription [6] – while more difficult in human cells, it would be interesting to ask whether relocation to the pore facilitates integration in a given gene. Finally, RanBP2 is known to have sumoylation activity [14], and it is of interest to investigate whether interactions with this pore protein lead to modification of the PIC with implications for integration site selection.

We imagine that the pathway to chromatin originating at the nuclear pore delivers the PIC to regions of the nucleus rich in gene-dense chromatin, and here LEDGF binds IN, tethering the complex to specific sites within this chromatin. However, it is possible that the pore provides direct access to LEDGF or to additional tethers. Studies in which both LEDGF and components of the pore are depleted to determine if effects on integration site selection are additive or epistatic may clarify these possibilities.

In Chapter 2, I presented data that suggested HIV-1 CA interactions with CypA and RanBP2 are antagonistic with respect to selection of chromatin: CypA appeared to decrease targeting to gene-dense chromatin while RanBP2 recovered some of the preference. Perhaps this balance ultimately enables optimal expression levels of the integrated HIV-1 genome. Alternatively, the modulation of integration
targeting by CypA may be a byproduct of a necessary interaction to protect the PIC from innate immune attacks in the cytoplasm until it arrives at the nuclear pore. Whether either interaction affected expression of the HIV-1 genome has yet to be investigated. Furthermore, it remains unknown whether CypA and RanBP2 bind CA competitively, and this information would color our interpretation.

How interactions with CypA might affect integration targeting downstream is no clearer than the role of the pore. The data presented in chapter 3 suggest that several different integration targeting pathways might diverge at the poorly defined step of viral uncoating and at nuclear import. Imaging of the mutant viruses discussed in this chapter in the cytoplasm and at the nucleus could shed light on these pathways and viral replication steps. Traditionally, microscopy studies of the PIC in an infected cell have been difficult, complicated by an inability to distinguish defective particles or particles on abortive pathways; however, recent advances in live-cell imaging of HIV-1 might make such studies possible [15].

A major goal of the work presented in Chapter 4 was to refine statistical models to better predict experimentally observed integration site distributions. Work to do this is still underway by our collaborator Dr. Charles Berry. However there are already signs that no combination of contributions of known preferred features will explain the data fully. A major hurdle following this study will be to identify the unknown contributors. One new avenue we plan to investigate is whether known sequence motifs recognized by various DNA-binding proteins are significantly enriched near preferred integration sites. Given our large dataset of
integration sites, we can consider orientation in this analysis, asking for the first
time whether components of the PIC interact with specific chromatin-bound factors
in an a-symmetrical fashion. Investigation for asymmetry of other genomic features,
such as CpG islands, with respect to proviral DNA could follow.

The wealth of annotations available along the human genome has grown rapidly, providing new features to consider. While we have analyzed correlations with histone methylations and acetylations, we have largely ignored other documented modifications such as phosphorylation and ubiquitination. Tan and colleagues recently reported 67 novel histone modifications [16], any of which might influence integration site selection in the genome. Intriguingly the distribution of one novel modification, lysine crotonylation (Kcr), is similar to that of HIV-1 integration sites with respect to transcription start sites of genes, and it's intensity correlates with gene expression [16]. Another new type of available genomic data comes from various chromosome conformation capture (3C) studies, that query sub-nuclear chromatin organization and interactions between multiple genomic loci [17]. It would be interesting to correlate this data with regions and sites in the genome that were more favorable than expected for integration. These “hot regions” might be compared to regions of equal length and predicted favorability that were not oversampled by the virus. Future work might also involve 3C-type studies in cells containing integrated proviruses.
Studies of HIV-1 mRNA populations

In Chapter 5, I presented a study of HIV-1 mRNA populations produced by an early passage viral isolate in a human osteosarcoma (HOS) cell line and in primary CD4+ T cells. I reported large variation between cell types, and temporal changes in HOS cells, and these differences were further examined in Chapter 6. Variation in mRNA populations between primary T cells from different human donors was significant but small, consistent with the importance of splicing regulators to host gene expression. In the above studies, the relative influences of splicing and mRNA stability on the observed ratios of HIV-1 transcripts were not investigated. While some groups have investigated decay rates of specific HIV-1 messages [18-21], a broad study of this kind has yet to be performed and would provide exciting insight into post-transcriptional regulation of HIV-1 gene expression. There is evidence to suggest that differences in stability may exist between cell types and viral strains [21], so it will be important to use clinically relevant cells and viruses for such a study. If we assume the differences in mRNA populations reported in Chapters 5 and 6 were due in part to splicing, it is of interest to ask whether the host-derived serine-arginine (SR)-rich proteins and heteronuclear ribonucleoproteins (hnRNPs) that regulate HIV-1 splicing in trans (Chapter 1 of this dissertation and [21]) differ in activity or abundance in ways that explain the data.

In Chapters 5 and 6, I reported variations in HIV-1 mRNA populations in which tat messages appear to be differentially affected. The splice acceptor used to make these messages is highly regulated, primarily in a repressive fashion [21].
Consistent with this, I showed in Chapter 5 that in T cells, *tat* mRNAs represent less than 5% of total messages. In Chapter 6, I presented studies showing that in HOS cells, this repression is alleviated with time, with Nocodazole treatment, or in the absence of Vpr. Thus, HIV-1 may be poised to rapidly induce Tat expression in response to a variety of stimuli, possibly involving regulation of the cell cycle and/or cytoskeleton. Understanding these stimuli and their importance in CD4+ T cells is a goal of future studies. One interesting experiment might involve studies of *cis-* and *trans-*regulation of *tat* in other complex retroviruses where latency may be more or less important. Additionally, these studies raise questions as to the role of the cell cycle and cytoskeleton in normal cellular splicing.

Given the array of differences between HOS and T cells and the suspected influences of cell-cycle progression and Vpr, studies in macrophages will be an important next step. Dowling and colleagues suggest that Tat levels may be temporally controlled in macrophages, *decreasing* with time as a result of virally induced modulation of host splicing factors. This may facilitate establishment of latent reservoirs in this cell type [18]. However, no in depth characterization of total HIV-1 mRNA populations have been done in these cells.

In Chapter 5, I described a new poorly conserved splice acceptor, A8c, downstream of the major splice site used in completely spliced transcripts, A7. A8c clustered with previously reported, conserved cryptic splice sites, and contributed to the abundant production of a class of small (~1kb) transcripts in cells infected with the HIV-1<sup>896</sup> strain. Preliminary work has suggested that transcripts in this
class are very rarely produced by several tested transmitted founder viruses, and also less abundantly produced by SIVmac239, which conserves A8c (unpublished). However, this class of transcripts was observed in PBMCs isolated from HIV-1 infected patients [22], therefore it would be of interest to test other clinical isolates of HIV-1.

It is unclear how frequently HIV-1 evolves abundantly used novel splice sites similar to A8c, and how these might affect the behavior of particular viral strains in which they occur. Such information may be clinically important, affecting selection of anti-retroviral drugs or other treatment options in patients harboring viruses with particular splice sites. HIV-1_{89.6} was isolated from the CSF, and exhibits unique behaviors, most strikingly the ability to form very large syncitia in macrophages [23]. Although it is presumed these syncitia are a function of the viral Env, it might be of interest to mutate A8c in HIV-1_{89.6} in order to ask whether splicing patterns contribute to the strain’s properties in-vitro. Importantly, the novel transcripts produced by HIV-1_{89.6} or other viral strains might also encode new immune epitopes or otherwise modulate immune responses, and such effects would be best studied in non-human primate models.

Finally, it has been suggested that HIV-1 might perturb cellular splicing machinery in a manner that benefits the virus, and this may be different in CD4+ T cells and macrophages [18,24,25]. The effects of HIV-1 infection on host expression in CD4+ T cells and on splicing of cellular mRNAs are currently the subjects of a study by Scott Sherrill-Mix, a graduate student in the Bushman lab.
Connections and Conclusions

It may not be coincidental that HIV-1 utilizes Transportin-3 to help determine its integration sites. Transportin-3 imports several SR proteins that promote splicing of the HIV-1 genome [12]. Most HIV-1 splicing is thought to be co-transcriptional, occurring at the site of integration [26]; therefore, it might be beneficial for the virus to traffic with these regulators and integrate in sub-nuclear regions where they may be concentrated. Preliminary studies of message populations in Transportin-3-depleted HOS-cells have not revealed noticeable changes in splicing populations (unpublished data), but detailed studies in relevant cell types have not been performed. Capsid mutations and knockdowns discussed in Chapters 2 and 3 as well as LEDGF fusion proteins that retarget HIV-1 integration [27] might all similarly be investigated to determine whether the site of integration affects mRNA populations. Alternatively, HIV-1 message RNA populations might be compared across multiple cell lines that each contain single or few HIV-1 integration sites.

The studies presented in this dissertation contribute to our understanding of several stages of the HIV-1 replication cycle, from uncoating of the viral core to gene expression of the virus and infected cell. They highlight the importance of host-virus interactions in the processes of HIV-1 integration and mRNA splicing, suggesting targets for new antiviral therapies, and illuminate vast networks of interconnected
processes important to normal cellular biology, with implications for a variety of human diseases.

References

APPENDIX

Appendix 1 - Guide to genomic heat map summaries of proviral distributions

The contents of this appendix have been published as a supplementary report in:
*equal contribution

General Description

We use heat maps to summarize the relationships of proviral distributions to genomic features. These maps were introduced in [1], which presents more background and examples of their uses. The goal of this Guide is to help users interpret comparisons summarized in the genomic features heat maps.

Generation of Tile Colors

Tile color indicates whether a chosen genomic feature is favored or disfavored by an integrating element, typically a retroviral vector, under the conditions studied. We determine this bias by asking how frequently that feature coincides with an integration site as compared to random sites in the genome using a receiver operating characteristic (ROC) curve area.

Generation of Matched Random Control sites: The calculation of ROC area is based on comparisons between true integration sites and computationally selected random sites. Because there is a bias in recovering integration sites
when using a restriction enzyme based method, k different matched random control (MRC) sites are selected for each individual integration site. Usually k will be at least 3 MRCs per integration site. Each MRC site lies the same distance from a restriction enzyme recognition sequence as the corresponding integration site but is otherwise randomly distributed in the human genome (described in [1,2]). For example, if the enzyme MseI is used to recover integration sites, and integration site A lies 120 bases from the nearest MseI site, all k MRCs selected for site A will also lie 120 bases from the nearest MseI site.

**Calculation of ROC area for genomic feature “J”:** The coincidence of genomic feature “J” with each integration site and matched random control site is measured (described for each feature below). Each integration site is then compared in a pair-wise fashion to its MRC sites, and a number is assigned indicating the relative rank of the integration site:

1 if the measurement of J is higher at the integration site than at the MRC site,
0 if the measurement of J is lower at the integration site than at the MRC site,
0.5 if the measurement of J is equal for the two sites.

All rank values thus calculated for a dataset of integration sites (all k rank values for all integration sites) are averaged to obtain the overall **ROC area** for the feature measured. An ROC area between 0 and 0.5 indicates the genomic feature occurs less frequently at/near integration sites than at/near random sites in the genome and is therefore disfavored. An ROC area between 0.5 and 1 indicates the genomic feature is enriched at integration sites. An ROC area of exactly 0.5 indicates that integration sites in the dataset are neither enriched nor depleted with respect to the feature of interest. The ROC area is converted to a color tile 201
according to the colorimetric scale at the bottom of the heat map. Positive associations (enrichment compared with random) are shown as increasing shades of red, negative associations (depletion compared with random) as increasing shades of blue, and no difference from random as white. Each tile represents a comparison to the randomly sampled controls for one genomic feature (row) in one experimental dataset (column).

Note that we do not present the magnitude of effect in terms of the original units of measurement. We simply ask whether the average integration site has a higher rank for a given type of feature than its k matched random control sites. The color indicates the average quantile of each integration site relative to its random controls. This removes skewing effects contributed by non-normal distributions of the data and also reduces the effect of a few data points with extreme values for a feature.

Statistics

Statistical tests to determine whether the ROC areas calculated are significantly different from one another or from 0.5 (matched random controls). These are described for each genomic feature below and in [1]. All the tests rely on the variance-covariance matrix of the relative ranks of the integration sites to construct Wald-type test statistics. For comparisons between integration site sets for specific genomic features, the Wald statistics are calculated and referred to the
Chi Square distribution to obtain p values as described in [3]. * = p<0.05, ** = p<0.01, *** = p<0.001

**Columns: Experimental Sets**

Each column of the heat map is a collection of colored tiles representing the preference of the vector for several genomic features under the condition tested (described below). The vector/experimental condition is listed at the top of the column.

**Rows: Genomic Features**

**Gene Boundaries:** Analysis of integration sites with respect to the transcription start and stop sites of genes as defined by indicated databases.

1. **In Gene, Unigene:** This tile indicates how frequently integration sites occur within genes in the Unigene database. The color of the tile can be interpreted as follows: Red indicates that the average integration site in the experimental set is more often found within a gene than matched random sites. Blue indicates that the average integration site is less often located within a gene than matched random sites. If the tile is white, the integration sites are no more or less likely to fall within a gene than matched random sites.

**ROC area calculation:** For each integration site-MRC site comparison, the integration site is scored:

- 1 if the integration site lies in a gene and the MRC site is outside a gene,
- 0 if the MRC site is in a gene and the integration site is not, and
0.5 if both sites fall within a gene or outside a gene.

ROC area is the average score across all comparisons (three per integration site) for all integration sites in the dataset.

Statistical Test for difference from Random (ROC area = 0.5): A one degree of freedom Wald test based on the variance mentioned above.

Significance: HIV-1 has a known preference for integration within active transcription units[4], so a set of integration sites of HIV-1 in wild type human cells will yield a red tile in this row.

2. In Gene, Refseq: Calculations and tile interpretations are identical to those for “In Gene, Unigene” with the exception that genes are defined by the Refseq database.

3. General Width: Indicates the relative width of the gene or intergenic space occupied by integration sites in the experimental set. The color of the tile can be interpreted as follows: Red indicates that the average space (gene width or intergenic width) occupied by integration sites in the set is larger than the average space occupied by matched random sites in the genome. Blue indicates that this average space is smaller for integration sites than for matched random sites. White indicates that this average space is the same for integration sites as it is for a collection of random sites in the genome.

ROC area Calculation: For an integration site that falls within a gene (Refseq), the width of the gene is measured in base pairs. Comparisons are made only to those of the three MRC sites that are also in genes. For
an integration site outside a gene, the width of the interval between the
nearest genes on either side of the site is measured. Comparisons are
made to the MRC sites outside genes. For each comparison the integration
site is scored:

1 if the gene or intergenic interval within which an integration site lies is larger
than that of the compared MRC site
0 if the interval is smaller for the integration site, and
0.5 if the intervals for the integration site and MRC site are matched in size.

The ROC area is the average score of all integration sites in the dataset.

Statistical Test for difference from Random Controls (ROC area = 0.5): A
one degree of freedom Wald test based on the variance mentioned above

Significance: The general width value provides one measure of gene
density, since gene dense regions are comprised of relatively short genes
and intergenic regions. This value also correlates negatively with gene
expression. Highly expressed genes tend to be shorter while less
expressed genes are longer. Similarly, actively transcribed genes tend to
cluster such that the intergenic width is shorter between expressed
genes, longer between less expressed genes. HIV prefers to integrate in
areas of the genome enriched for actively transcribed genes [4], therefore
this tile will be blue for HIV-1 in wild type cells.

4. Gene Width: Displays the relative width of the gene occupied by integration
sites in the experimental set. Red indicates that when integration occurs in
genes, there is a bias for larger genes. Blue, that integration is favored in
shorter genes. The tile is white if integration shows no bias for either short or
long genes. Calculations and interpretations are similar to those for "General width"; however, for this tile, only those integration site-MRC site pairs in which both sites fall within genes are considered.

5. **Distance to Start**: Indicates whether integration is preferred near transcription start sites (TSS). A red tile means that integration sites in the experimental set are farther from TSSs than matched random sites in the genome and that gene start sites are disfavored for integration. Blue indicates that integration sites lie closer to gene starts than do random sites (gene start sites are favored). If the tile is white, TSSs are neither favored nor disfavored for integration. Note that in contrast to most features in the heat map, gene start sites are **favored** when this tile is blue, not red, indicating the shorter distance to start sites.

**ROC area Calculation**: For sites that fall within genes, we measure the distance in base pairs to the TSS of that gene. For sites outside of genes, we measure the distance to the nearest transcription start site (according to RefSeq). An integration site is scored:

1 if the distance to the nearest TSS is larger for the integration site than for the compared MRC site
0 if the distance to the nearest TSS is smaller for the integration site
0.5 if the integration site and MRC site are equidistant from the nearest TSS.

The ROC area is the average score of all integration sites in the dataset.

**Statistical Test for difference from Random Controls (ROC area = 0.5)**: A one degree of freedom Wald test based on the variance mentioned above
**Example/Significance:** MLV and other gammaretroviruses show a strong preference for integration within gene promoters [5] and therefore, this tile would appear blue for an MLV infection of wild type cells. This makes such viruses relatively dangerous for use as gene therapy vectors, as integration in promoters may alter the control of important host genes [6-8] [6-10]. By contrast, lentiviruses show little preference for gene starts and this tile appears white for HIV-1 infection of wild type cells [2,4]. This measurement is also influenced by the preference of the vector studied for integration within genes as well as width of gene or intergenic space occupied.

6. **Distance to Boundary:** Indicates whether integration is preferred near the boundaries (transcription start or end) of genes. As with “Distance to Start” blue indicates that gene boundaries are favored for integration. Red indicates that integration disfavors gene boundaries and white indicates that the virus has no preference for gene boundaries under the conditions studied.

**ROC area Calculation:** Calculations are similar to those for “Distance to Start” except that for all sites, the distance in base pairs to the nearest gene boundary (start or end) is considered.

**Statistical Test for difference from Random Controls (ROC area = 0.5):** A one degree of freedom Wald test based on the variance mentioned above

**Significance:** This measure is influenced by the preference of the vector studied for integration within genes, the width of gene or intergenic space
occupied and the preference for transcription start sites (described above).

7. <50 kb from Oncogene: Indicates how frequently integration sites occur within 50 kb of an oncogene (UCSC hg18 goldenpath database for genes, compared to the allOnco cancer-related gene list at http://microb230.med.upenn.edu/protocols/cancergenes.html). Red indicates that integration sites occur more frequently than random sites near oncogenes. Blue indicates that integration near oncogenes is disfavored. If the tile is white, the integration sites are no more or less likely to fall near an oncogene than matched random sites.

**ROC area calculation:** An integration site is scored:

1 if the integration site falls < 50kb from an oncogene and the compared MRC site is ≥50 kb from an oncogene,
0 if the MRC site is <50 kb from an oncogene and integration site is not, and
0.5 if both sites lie within or outside 50 kb of an oncogene.

ROC area is the average score for all integration sites in the dataset.

**Statistical Test for difference from Random Controls (ROC area = 0.5):** A one degree of freedom Wald test based on the variance mentioned above

**Significance:** Proximity of integration sites to oncogenes is important to consider when choosing vectors to use in gene therapy applications.

**DNase Sites:** Analysis of integration sites with respect to the location of DNase hypersensitive sites in the genome. We ask whether integration sites fall more or less frequently within the indicated distance of a DNase hypersensitive site (<1Mb,
<100kb and <10kb) than would be expected for a random distribution (location of DNase sites from UCSC hg18 goldenpath database). Red indicates a preference for integration near a DNase hypersensitive site. Blue indicates that integration is disfavored near DNase hypersensitive sites. If the tile is white, the integration sites are no more or less likely to fall within the indicated distance of a DNase hypersensitive site than matched random sites.

**ROC area Calculation (eq <1Mb):** An integration site is scored:

1 if the integration site falls within 1 Mb of a DNase hypersensitive site and the compared MRC site does not,
0 if the MRC site is <1 Mb from a DNase site and the integration site is not,
0.5 if both sites lie within or outside 1 Mb of a DNase hypersensitive site.

ROC area is the average score for all integration sites in the dataset.

**Statistical Test for difference from Random Controls (ROC area = 0.5):** A one degree of freedom Wald test based on the variance mentioned above

**Significance:** DNase hypersensitive sites are exposed in more active chromatin and are protected in heterochromatin. Constitutively accessible DNase hypersensitive sites are surrogate markers for open active areas of chromatin. HIV-1 prefers to integrate within actively transcribed regions, and therefore in regions rich in DNase sites. This preference appears most strongly at larger window sizes and serves as a marker for gene dense regions. HIV disfavors integration near DNaseI sites at very short distances, paralleling the disfavoring of CpG islands and gene start sites ([2,11]. MLV, in contrast, strongly favors integration very near DNaseI hypersensitive sites [11].
**CpG Islands**: Analysis of integration sites with respect to the location of CpG islands within the genome (UCSC hg18 goldenpath database).

1. **CpG Density, 1Mb, 100kb, and 10kb**: We ask whether integration sites fall within regions more or less dense in CpG islands than would be expected for a random site in the genome. To do so we consider the density of CpG islands within the indicated genomic intervals surrounding each integration site. A red tile indicates that integration occurs in regions enriched for CpG islands. Blue indicates that integration favors regions poor in CpG islands compared to random. White indicates that integration sites show no bias for or against regions enriched in CpG islands.

   **ROC area Calculation (e.g. CpG Density, 1Mb)**: We count the number of CpG islands falling in a 1Mb window surrounding each integration and MRC site (ie within 500 kb of a site). An integration site is scored:

   1 if more CpG islands lie in the 1Mb surrounding the integration site than in the same interval surrounding the compared MRC site,
   0 if fewer CpG islands lie in the genomic interval around the integration site,
   0.5 if same number of CpG islands fall in the interval surrounding each site.

   ROC area is the average score for all integration sites in the dataset.

2. **<5kb and <1kb**: Indicates whether integration sites fall more or less frequently within the indicated distance of a CpG island than would be expected for a random distribution (location of CpG islands from UCSC hg18 goldenpath database). Red indicates a preference for integration near a CpG
island. Blue indicates that integration is disfavored near CpG islands. If the tile is white, the integration sites are no more or less likely to fall near CpG islands than matched random sites.

ROC area Calculation (e.g. <5kb): An integration site is scored:

1 if the integration site falls <5kb from a CpG island and the compared MRC site does not,
0 if the MRC site is <5kb from a CpG island and the integration site is not,
0.5 if both sites lie within or outside 1 Mb of a CpG island.

ROC area is the average score for all integration sites in the dataset.

Statistical Test for difference from Random Controls (ROC area = 0.5): A one degree of freedom Wald test based on the variance mentioned above

Significance: CpG islands are enriched in/near gene promoters, especially those of housekeeping genes, and are otherwise rare within the genome due to methylation and deamination of the cytosine. They are therefore markers of promoters when small genomic windows surrounding CpG islands are considered, and of gene dense regions when larger windows are considered. That is if a site is very close to a CpG island it is likely to be close to a gene promoter, and if a site is in a broad region enriched for CpG islands, it is likely to be in a more gene dense area of the genome.

Gene Density: Analysis of integration sites with respect to the local density of genes. We consider the indicated genomic interval surrounding each integration site (1Mb, 100kb, or 10kb) and ask whether integration sites fall within regions more or less dense in genes than would be expected for a random site in the genome. A red tile
indicates that integration occurs in regions enriched for genes. Blue indicates that integration favors regions poor in genes compared to random. White indicates that integration sites show no bias for or against gene dense regions.

**ROC area Calculation (e.g. Density 1Mb):** We count the number of genes (RefSeq) falling in a 1Mb window surrounding each integration and matched random control site (i.e. within 500 kb of a site). An integration site is scored:

1 if more genes lie in this interval for the integration site than for the compared MRC site,
0 if fewer genes lie in the interval around the integration site, and
0.5 if same number of genes fall in a 1Mb interval surrounding each site.

ROC area is the average score for all integration sites in the dataset.

**Statistical Test for difference from Random Controls (ROC area = 0.5):** A one degree of freedom Wald test based on the variance mentioned above

**Significance/Example:** HIV has a known preference for integration in regions of the chromosome that are enriched for genes. This may be because it prefers to integrate within active transcription units [4]. However it may also be independently guided to regions of chromatin enriched for genes. In either case, this preference is hypothesized to afford HIV better access to transcription factors.

**Expression Intensity:** Analysis of integration sites with respect to the local density of sets of genes on an Affymetrix Gene Chip. We consider density of total genes as well as density of highly expressed genes. For this analysis we use microarray data measuring relative expression of genes in the cell type used in the study. We note
that manipulations to cells in certain experiments may change gene expression profiles, and this must be considered in interpretation of the data.

1. **All Genes Density, 1Mb**: Calculations and tile interpretations are identical to those for “Density, 1Mb” in the Gene Density section with the exception that the “genes” counted are loci identified by expression probe sets on the relevant Affymetrix GeneChip.

2. **Top ½ Expression, 1Mb**: Calculations and tile interpretations are identical to those for “All Genes Density, 1Mb” with the exception that only the top ½ most expressed loci are counted.

3. **Top 1/16th Expression, 1Mb**: Calculations and tile interpretations are identical to those for “All Genes Density, 1Mb” with the exception that only the top 1/16th most expressed loci are counted.

*Statistical Test for difference from Random Controls (ROC area = 0.5)*: A one degree of freedom Wald test based on the variance mentioned above

*Significance*: This analysis allows us to consider whether the preference for total genes is the same as the preference for highly expressed genes. This would be the case if the tile colors are the same for all three measures. If highly expressed genes are preferred more than are genes in general, the “Top 1/16th Expression” and “Top 1/2 expression” tiles may be red shifted from the “All Genes Density” tile.
**GC Content**: Analysis of integration sites with respect to the local GC content. A red tile indicates that for the genomic interval considered (see below for discussion of window size) integration sites lie in regions that are GC-rich compared to random sites in the genome. Blue indicates that integration is disfavored in GC rich regions (favored in AT rich regions). White indicates that the vector shows no bias for GC content at the genomic interval considered.

**ROC area Calculation**: GC content is measured within the indicated interval surrounding each integration and matched random site. An integration site is ranked:

1 if the defined region surrounding an integration site is more GC rich than that surrounding the compared MRC site
0 if the integration site lies in a less GC rich region than the MRC site, and
0.5 if the integration and MRC sites are located in equally GC rich regions.

ROC area is the average rank for all integration sites in the dataset.

**Statistical Test for difference from Random Controls (ROC area = 0.5)**: A one degree of freedom Wald test based on the variance mentioned above

**Significance**: GC content is positively correlated with genes and therefore when considering broad windows surrounding integration sites, a high GC content may indicate a preference for genes/gene dense regions. However, due to local sequence constraints that affect the choice of integration sites of many viruses, small windows surrounding integration sites may show different effects. For example, HIV prefers to integrate in gene dense regions and, predictably, GC content is high if measured in large windows surrounding HIV-1 integration sites in wild type cells. However if we consider only small windows
surrounding integration sites, we find that HIV-1 prefers relatively AT-rich loci within those large GC rich regions. This is likely to be secondary to HIV-1’s preference for integration on nucleosomes which position in relatively AT rich DNA [12]. The AT-hooks of LEDGF-p75, a known tether for HIV integrase, may also contribute to this effect [13]. For this reason, we display ROC area tiles for GC content in several genomic intervals.

References


Appendix 2 – Curriculum Vitae

Education

University of Pennsylvania, Philadelphia, PA 2006 - present
Candidate for M.D.-Ph.D. combined degree
Patterns of HIV integration and splicing: windows on mechanism.
Thesis Advisor: Dr. Frederic D. Bushman
Expected completion of dual degree: May 2014

Brown University, Providence, RI 2001 - 2005
Bachelor of Science in Biochemistry
Magna Cum Laude, Phi Beta Kappa
Mechanisms of mouse resistance to Toxoplasma gondii: role of IFN-inducible GTPase.
Thesis Advisor: Dr. George S. Yap

Laboratory Research Experience

DePuy Biologics (Johnson & Johnson), Raynham, MA 2005 - 2006
Research intern: In vitro modeling of degenerative disc disease.

Albert Einstein College of Medicine, Bronx, New York Summer 2003
Research on lipid antigen recognition, in the laboratory of Dr. Steven Porceli.

Regeneron Pharmaceuticals, Inc., Tarrytown, NY 1999 - 2001
Research on a novel gene using transgenic technology in the laboratory of Dr. Thomas DeChiara.

Professional Activities

Reviewer: Journal of Virology 2011-Present
Teaching: Rotation student supervisor 2010-2011
Teaching assistant in genetics, Brown University 2002
Membership: Women’s M.D.-Ph.D. Student association, U. Penn. 2010-present

Honors and Awards

Richard K. Root Prize for Infectious Disease Research 2012
James S. Porterfield Prize in International Virology 2012
NIH T32 Training Grant 2010-2011
Maria L. Caleel Memorial Award for Academic Excellence 2005
Howard Hughes Undergraduate Summer Research Fellowship 2004
Brown University Junior Prize in Biochemistry 2004
Merck Prize for Outstanding Performance in Organic Chemistry 2003
Publications


Conference Talks


Conference Posters
