EXPRESSION, LATENCY AND SPLICING DURING INFECTION WITH HIV

Scott Sherrill-Mix

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

Frederic D. Bushman, Ph.D., Professor of Microbiology

Graduate Group Chairperson

Li-San Wang, Ph.D., Associate Professor of Pathology and Laboratory Medicine

Dissertation Committee:

Nancy Zhang, Ph.D. Associate Professor of Statistics

Yoseph Barash, Ph.D., Assistant Professor of Genetics

Kristen W. Lynch, Ph.D., Professor of Biochemistry and Biophysics

Michael Malim, Ph.D., Professor of Infectious Diseases, King's College London

EXPRESSION, LATENCY AND SPLICING DURING INFECTION WITH HIV

Dedicated to William Maurer, Gayle Maurer & Michele Sherrill-Mix

ACKNOWLEDGEMENTS

I would like to thank

Rick

Bushman lab Chris wetlab

Hannah Chervitz and Tiffany Barlow

GCB

committee

collaborators

Xiaofen and Otto

HINT grant and GCB training grant

. . .

ABSTRACT

EXPRESSION, LATENCY AND SPLICING DURING INFECTION WITH HIV

Scott Sherrill-Mix

Frederic D. Bushman, Ph.D.

No more than 350 words. It is normally a single paragraph, consists of four parts: the statement of the problem; the procedure and methods used to investigate the problem; the results of the investigation; and the conclusions. The abstract is published online by ProQuest in “Dissertation Abstracts International”, providing information to interested readers about the general content of the dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

1.1 Impact of HIV

In 1981, physicians began to notice a mysterious increase in the occurrences, often clustered in men who had sex with men or intravenous drug users, of Kaposi's sarcoma and pneumocystis pneumonia [1–6].

Kaposi's sarcoma had been a rare cancer in the US found largely in elderly men with Jewish or Mediterranean ancestry [7]. Kaposi's sarcoma had also been seen in immunocompromised individuals [8–10] and there were suggestions that it was a virus-associated cancer [11] although the causative human herpesvirus 8 would not be discovered for another decade [12, 13].

Pneumocystis pneumonia was known to be caused by infection of the alveoli with the yeast-like fungus *Pneumocystis jirovecii* [14, 15], previously known as *Pneumocystis carinii* [16]. Pneumocystis pneumonia was almost exclusively seen patients with suppressed immune systems or immune disorders and rarely if ever in immunocompetent individuals [15].

The mechanism for this spike of these opportunistic infection was clarified when researchers found severe T cells depletion and decreases in cellular immunity in these patients [4–6, 17, 18] but the underlying cause remained unclear and eventually termed acquired immune deficiency syndromek. Potential transmissions by transfusion [19, 20], injection drug use [4, 18, 21] and both homosexual [17, 22] and heterosexual [18, 23] contact pointed towards a transmissible agents [[double transmission]]. A virus named lymphadenopathy-associated virus or human T-lymphotropic viruses III and later renamed human immunodefficiency virus was isolated from patient samples in 1983 [24**?** , 25] and soon detected in most immunodeficient patients [26–28]. The virus was sequenced in 1985 [29].

History of the virus and infected population

Origin in chimpanzee [30]

1.2 The HIV virus

1.3 Integration and latency

1.4 Host cell interactions

1.5 HIV splicing

Driven by a strong selective pressure for genome compactness [31–33], HIV and other lentiviruses subvert host cell alternative splicing pathways to allow tight packing of their genetic information. Through weak splice sites and overlapping reading frames (Figure **??**), the virus manages to produce precise quantities of at least nine proteins and polyproteins from its single transcription start site and less than 10 kb genome [34].

As such an integral part of the virus life cycle[35, 36], alteration of splicing poses a tempting therapeutic target. Inhibition of cellular splicing factors reduces viral reproduction in many genome-wide siRNA screens [37–39] and several members of the spliceosome interact with viral proteins in affinity pulldowns [40]. Open reading frames in uncharacterized transcripts appear to produce epitopes useful for vaccine development [41]. Potential treatments altering viral splicing through small molecule inhibitors [42, 43] and gene therapy [44, 45] have restricted viral replication *in vitro*. However without methods to quantify viral splicing or a thorough quantification of splicing under varying conditions, the development of such treatments remains limited.

Viral proteins also interact with components of the cellular splicing complex [40, 46, 47]. These interactions have been reported to change splicing in viral[47–49] and cellular transcripts [50, 51] and raise the possibility that the virus has evolved to alter host

splicing. A genome-wide study of changes in cellular splicing during HIV infection would greatly clarify this hypothesis but no such study has been performed.

Alternative splicing, the differential inclusion of exons and removal of introns from primary mRNA transcripts, allows rapid evolution of protein segments [52–54] and drastic increases in the number of proteins generated by a single DNA sequence [55]. Many viruses subvert the splicing machinery of their eukaryotic hosts to modify their viral mRNA [56].

In particular, it has previously been reported that HIV utilizes alternative splicing to generate more than 40 mRNA transcripts encoding at least 9 proteins and polyproteins from a genome smaller than 10kb [57]. A specific progression of viral transcripts appear in the cytoplasm of the host cell as infection progresses allowing a shift from regulatory protein production in early infection into virion production in late infection [35, 36, 58]. Because HIV has only a single transcription start site, these transcriptional changes are driven by alternative splicing [34].

Although it plays such an essential role for the virus, only a single detailed census of viral splicing has been reported [57]. Due to limitations in technology, this study was limited to only the most abundant transcripts in lab-adapted HIV strains in cell culture [57]. Yet rare transcripts may play an important role in immune response [41] and encode unknown proteins [59]; lab adapted HIV can differ markedly from viruses actually found in patients [60]; cell cultures often do not reflect *in vivo* conditions [61]; and splicing can vary between humans [62, 63] and cell types [64, 65]. Without a fuller characterization of transcripts under these relevant conditions, many aspects of viral splicing will remain poorly understood.

Alternative splicing may also play an unappreciated role in HIV-host interactions. Viral proteins interact with the splicing complex [40, 46, 47] and alter splicing of some cellular transcripts [50, 51]. Yet, although infection has been shown to cause

genome-wide changes in the expression of cellular genes [66–70], no genome-wide study of cellular alternative splicing during HIV infection has ever been reported. Such a genome-wide study of splicing changes might reveal a distortion of diverse cellular splicing which is adaptively advantageous to the virus.

Current sequencing advancements allow a much broader and deeper investigation of viral splicing. Targeted amplification with RainDance droplet PCR offer the potential to reduce size bias inherent in bulk PCR [71]. RNA-seq with Illumina sequencing allows extremely deep sequencing of cellular and viral transcripts with billions of bases of short read sequence [72, 73]. Single molecule sequencing with Pacific Biosciences provides reads approaching 20,000 bases [74, 75] that could characterize entire viral transcripts in one continuous read. By combining these technologies, viral and cellular transcripts could be interrogated to an unprecedented level.

A better understanding of viral splicing and viral effects on host splicing may bring therapeutic benefits. siRNA inhibition of splicing factors reduces HIV replication in many genome-wide screens [37–39]. Alteration of viral splicing through small molecule inhibitor of SR protein kinases[42] and Splicing Factor 2 [43], shRNA against spliceosomal U7 snRNP [44] and expression of modified spliceosomal U1 snRNP [45] show treatment potential *in vitro*. In addition, rare uncharacterized HIV transcripts and their encoded proteins appear to produce potent immune response in HIV patients [41] thus offering potential targets for vaccine development. Yet without methods to characterize viral RNA and measure the effects of treatments on viral splicing, further development is inhibited.

Inclusion and exclusion of a particular stretch of RNA into an mRNA is determined by a balance of RNA secondary structure [76–78], chromatin structure [79], nucleosome positioning [80], histone marks [81], previous splicings [82], order of intron removal [83, 84] and enhancers [85] and suppressors [86] that bind specific motifs [87]. Together these factors create a precise controllable splicing code [65, 88, 89].

In HIV, splicing occurs between at least four splice donors and eight splice acceptors [34]. Two splice donors, D1 and D4, are relatively strong while the remaining donors and all acceptors are fairly weak [90]. Several exonic splicing silencers [91, 92] and exon splicing enhancers [93, 94] and a single intronic splicing silencer [95] in the viral genome interact with many human splicing factors, including hnRNPs A1 [92, 95] H, F, 2H9, and A2 [77] and SR proteins SRp40[93, 96], SRp75 [96], ASF/SF2 [93] and SC35 [77], to alter viral splicing [34, 97].

Several viral proteins affect mRNA abundances. Rev causes export of unspliced viral mRNA that would otherwise be trapped in the nucleus [98] to be exported [56, 99] and may also interact with splicing factors to alter viral splicing [46]. The HIV protein Tat is best known for its transactivation of viral transcription [100, 101] and triggering apoptosis in uninfected cells [102, 103] but Tat also appears to independently affect alternative splicing of viral transcripts[47–49, 104]. Viral protein Vpr is known to cause cell cycle arrest [105] and mediate nuclear import of the viral preintegration complex [106]. Vpr also appears to alter alternative splicing of some cellular transcripts [50, 51] and interact with the SMN complex [40], which assembles spliceosomal snRNP [107]. Although all three of these proteins modify viral splicing, whether they also cause widespread alterations in cellular splicing is unknown.

Despite the critical role alternative splicing plays in viral replication, no genome-wide studies of lentiviral effects on cellular splicing or detailed censuses of viral alternative splicing in biologically relevant settings have been published.

RNA-seq offers a much broader view of alternative splicing than previously possible [108, 109] but Illumina sequencing has not yet been applied to the study of differential splicing in host RNA of HIV-infected cells. There have been many studies of cellular expression using microarrays [66–69, 104] and Sage [110, 111] but only a single study using Illumina RNA-seq and alternative splicing changes were not reported [70]. Thus a potentially significant aspect of HIV-host interactions remains unknown.

The most extensive survey of HIV transcripts to date was published in 1993[57]. Technology at the time necessitated the use of Northern blots and RNA protection assays [57] which can not distinguish multiple similarly sized transcripts or detect rare transcripts. This study also focused on a single lab adapted HIV_{NL4-3} strain in HeLa cell culture.

Many previous studies of viral splicing have used lab-adapted strains of HIV which often differ from patient isolates [60] in cell cultures which often differ from primary cells [61]. Selection under cell culture conditions may quickly alter splicing patterns to down regulate proteins unneeded *in vitro*. Characterization of alternative splicing in biologically relevant cell types infected with clinical isolates of HIV are sorely needed.

1.6 RNA detection

First HIV antibody test [27, 28]

CHAPTER 2 : Gene activity in primary T cells infected with $HIV_{89.6}$: intron retention and induction of distinctive genomic repeats

2.1 Abstract

Background: HIV infection has been reported to alter cellular gene activity, but published studies have commonly assayed transformed cell lines and lab-adapted HIV strains, yielding inconsistent results. Here we carried out a deep RNA-Seq analysis of primary human T cells infected with the low passage HIV isolate $HIV_{89.6}$.

Results: Seventeen percent of cellular genes showed altered activity 48 hours after infection. In a meta-analysis including four other studies, our data differed from studies of transcription after HIV infection of cell lines but showed more parallels with infections of primary cells. We found a global trend toward retention of introns after infection, suggestive of a novel cellular response to infection. $HIV_{89.6}$ infection was also associated with activation of human endogenous retroviruses (HERVs) and several retrotransposons, of interest as possible novel antigens that could serve as vaccine targets. The most highly activated group of HERVs was a subset of the ERV-9, a group not reported previously to be induced by HIV. Analysis showed that activation was associated with a particular variant of an ERV-9 long terminal repeat that contains an indel near the U3-R border. These data also allowed quantification of > 70 splice forms of the $HIV_{89.6}$ RNA and specified the main types of chimeric $HIV_{89.6}$-host RNAs. Comparison to 147,281 integration site sequences from the same infected cells allowed quantification of authentic versus artifactual chimeric reads (0.1% of the total), showing that $5'$ read-in, splicing out of $HIV_{89.6}$ from the D4 donor and $3'$ read-through were the most common $HIV_{89.6}$-host cell chimeric RNA forms.

Conclusions: Analysis of RNA abundance after infection of primary T cells with the low passage $HIV_{89.6}$ isolate disclosed multiple novel features of HIV-host interactions, notably intron retention and induction of transcription of distinctive retrotransposons

and endogenous retroviruses.

2.2 Background

HIV replication requires integration of a cDNA copy of the viral RNA genome into cellular chromosomes, followed by transcription and splicing to yield viral mRNA. Alternative splicing allows the small 9.1kb HIV genome to generate at least 108 mRNA transcripts encoding at least 9 proteins and polyproteins [29, 34, 57, 112–114]. During replication, HIV also reprograms cellular transcription and splicing. For example, the virus-encoded Vpr protein arrests the cell cycle [105, 115–117] and the viral Tat protein binds to P-TEFb and alters transcript at the HIV promoter and some cellular promoters [118–123].

Multiple studies suggest that cells detect HIV infection and respond by inducing interferon-regulated, apoptotic and stress response pathways [70, 124–131]. Several studies have also suggested that HIV infection disrupts normal cellular splicing pathways [131, 132]. However, results have varied with many experimental parameters, including target cell type, HIV isolate and the duration of infection. Many of the published studies focused on infections with lab-adapted HIV strains in transformed cell lines [68, 70, 111, 124, 131, 133], and so results may not be fully reflective of infections in patients.

In this study, we sought to generate data more resembling HIV replication in patients by analyzing transcriptional responses after infection of primary T cells with $HIV_{89.6}$, a low passage patient isolate [134]. This represents a continuation of a long term effort to understand HIV-host cell interactions at the transcriptional level that began with analysis of transcription by $HIV_{89.6}$ in primary T cells using Pacific Biosciences long read single molecule sequencing [114]. Our strategy here was to analyze a single time after infection in depth, analyzing over 1 billion sequence reads from $HIV_{89.6}$ infected and uninfected host cells. These data were then combined with 147,281 unique

integration site sequences from the same infections and the Pacific Biosciences data on $HIV_{89.6}$ transcription to 1) elucidate effects of HIV infection on host cell mRNA abundances and splicing, 2) characterize viral message structure in detail and 3) probe the nature of the chimeras formed between host cell and viral RNAs.

2.3 Methods

2.3.1 Cell culture and viral infections

$HIV_{89.6}$ stocks were generated by the University of Pennsylvania Center for Aids Research. 293T cells were transfected with a plasmid encoding an $HIV_{89.6}$ provirus, and harvested virus was passaged in SupT1 cells once. Viral stocks were quantified by measuring 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 single healthy male donor (ND365) using the RosetteSep Human $CD4^+$ T Cell Enrichment Cocktail (StemCell Technologies).

T cells were stimulated for 3 days at 0.5×10^6 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) + 5μg/mL PHA-L (Sigma-Aldrich)). Cells were infected in triplicate and mock infections were performed in duplicate. For each infection, 6.6×10^6 cells were mixed with 1.32 μg $HIV_{89.6}$ in a total volume of 2.25 mL. Infection mixtures was split into 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. 1×10^6 cells were harvested for flow cytometry, and 6×10^6 cells were pelleted following two washes in PBS for nucleic acid extraction. Genomic DNA and total RNA were isolated from 6×10^6 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. RNA samples were treated with DNase prior to elution in 40 μL water.

2.3.2 Analysis of HIV$_{89.6}$ integration sites in primary T cells

Integration site sequences were determined for DNA fractions from the above infections after ligation mediated PCR [135]. A total of 147,281 unique integration site sequences were determined. An analysis of integration site distributions for these samples was reported in Berry et al. [135].

2.3.3 mRNA sequencing

Messenger RNA was isolated and amplified from purified total cellular RNA (3 μL or approximately 9 μg from each uninfected sample, 25 μL or approximately 3 μg from each infected sample) using the Illumina TruSeq RNA sample preparation kit according to manufacturer's protocol. SuperScript III (Invitrogen) was used for reverse transcription. Each sample was tagged with a separate barcode and sequenced on an Illumina HiSeq 2000 using 100-bp paired-end chemistry.

2.3.4 Flow cytometry

To assess percent infected cells, 1×10^6 cells per infection were stained for flow cytometry. All staining 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).

2.3.5 Analysis

Reads were aligned to the human genome using a combination of BLAT [136] and Bowtie [137] through the Rum pipeline [138]. Estimates of fragments per kilobase of transcript per million mapped reads and changes in expression for cellular genes were calculated by Cufflinks [108]. Reads found to contain sequence similar to the HIV genome using a suffix tree algorithm were aligned against the $HIV_{89.6}$ genome using BLAT [136]. All statistical analyses were performed in R 3.1.2 [139]. RNA-Seq reads from Chang et al. [70] were downloaded from the Sequence Read Archive (SRP013224) and aligned using the Rum pipeline.

Gene lists were obtained from the supplementary materials of four other studies of differential gene expression during HIV infection [70, 111, 129, 140]. We called genes differentially expressed in Li et al. [140] if they had a reported $p < 0.01$ or in Lefebvre et al. [111], Chang et al. [70] and Imbeault et al. [129] if they had an adjusted $p < 0.05$. We called genes as differentially expressed in our own study if the adjusted $p < 0.01$. For the comparison of differentially expressed genes regardless of direction in figure 2.1 (below the diagonal), it was unclear exactly how many genes were studied in each study so we assumed a background of the 14,192 genes (the number of genes which could be tested for significance in our data).

We obtained transcriptional profiles comparing immune cell subsets from the Molecular Signatures Database [141]. MSigDB set names from the MSigDB used in Figure 2.2A were GSE10325 LUPUS CD4 TCELL VS LUPUS BCELL, GSE10325 CD4 TCELL VS MYELOID, GSE10325 CD4 TCELL VS BCELL, GSE10325 LUPUS CD4 TCELL VS LUPUS MYELOID, GSE3982 MEMORY CD4 TCELL VS TH1, GSE22886 CD4 TCELL VS BCELL NAIVE, GSE11057 CD4 CENT MEM VS PBMC, GSE11057 CD4 EFF MEM VS PBMC, GSE3982 MEMORY CD4 TCELL VS TH2 and GSE11057 PBMC VS MEM CD4 TCELL and in Figure 2.2B were GSE36476 CTRL VS TSST ACT 72H MEMORY CD4 TCELL OLD, GSE10325 CD4 TCELL VS LUPUS CD4 TCELL, GSE22886 NAIVE

CD4 TCELL VS 12H ACT TH1, GSE3982 CENT MEMORY CD4 TCELL VS TH1, GSE17974 CTRL VS ACT IL4 AND ANTI IL12 48H CD4 TCELL, GSE24634 IL4 VS CTRL TREATED NAIVE CD4 TCELL DAY5, GSE24634 NAIVE CD4 TCELL VS DAY10 IL4 CONV TREG, GSE1460 CD4 THYMOCYTE VS THYMIC STROMAL CELL and GSE1460 INTRATHYMIC T PROGENITOR VS NAIVE CD4 TCELL ADULT BLOOD.

We downloaded the RepeatMasker track from the UCSC genome browser [142] and used the SAMtools library [143] to assign reads to the repeat regions. HERV-K age estimates were obtained from the supplementary materials of Subramanian et al. [144].

We used a Bayesian estimate of the ratio of expression in uninfected and HIV infected samples to account for sampling effort and differing expression in genomic regions. We modeled the observed counts as a binomial distribution with a flat beta prior ($\alpha = 1, \beta = 1$) separately for uninfected and infected samples. We then Monte Carlo sampled the two posterior distribution to estimate the posterior distribution of the ratio. For introns, the number of binomial successes was set to the number of reads mapped to the intron and the number of trials was the total number of reads observed in the genes overlapping that intron. For repeat regions, the number of binomial successes was set to the number of reads mapped to that region and the number of trials was the total number of reads mapped to the human genome.

To estimate determinants of LTR12C expression, we fit a logistic regression for which LTR12C increased in expression with $\text{HIV}_{89.6}$ infection (95% Bayesian credible interval > 1) on to characteristics of the LTR12C regions. We extracted all the LTR12C regions from the human genome and determined the U3-R boundary using a ends free alignment of the previously reported U3-R border [145–149] against the sequences. Regions less than 1,000 bases long were discarded. Previous studies disagreed about the location of the LTR12C transcription start site and it appears that transcription may start in several places [146, 147]. We took the $5'$ most site that had agreement between

studies (transcription starting with TGGCAACCC). We split the sequences into short, medium and long length classes based on an indel about 70 bases upstream from the transcription start site. For each length class, we generated a consensus sequence and counted the Levenshtein edit distance between the consensuses and each corresponding sequence. We also counted the number of NFY motifs (CCAAT or ATTGG), MZF1 motifs (GTGGGGA) and GATA2 motifs (GATA or TATC) in the entire U3 region or checked in any of the three motifs was present in the 150 bases upstream of the TSS. A final regression model was selected using stepwise regression with an AIC cutoff of 5. For display, the LTR12C sequences were aligned with MUSCLE [150].

The abundance of the HIV RNA size classes was estimated as described in Additional File 5. These estimates were then multiplied by the within size class proportions estimated by Ocwieja et al. [114] using PacBio sequencing of $HIV_{89.6}$ to yield proportions over 78 measured $HIV_{89.6}$ RNAs.

2.4 Results

2.4.1 Infections studied

$HIV_{89.6}$, a clade B primary clinical isolate [134], was used to infect primary $CD4^+$ T cells from a single human donor in three replicate infections. For comparison, two additional replicates from the same donor were mock infected. Samples were harvested after 48 hours of infection, which allowed for widespread infection in the primary T cell cultures, though some cells may be infected secondarily by viruses produced in the first round. Thus cultures probably were not tightly synchronized but did have extensive representation of infected primary T cells. From these samples, we obtained 1,161,705,678 101-bp reads from primary $CD4^+$ T cells from a single donor; 1,021,207,853 were mapped to the human genome and 24,783,844 to the $HIV_{89.6}$ provirus (Table 2.1). Below we first discuss the influence of infection on cellular gene activity and RNA splicing, then analyze HIV RNAs and lastly analyze chimeras formed between HIV and cellular

Sample	Infection rate (%)	Reads	Human reads	HIV reads	% HIV	% HIV in infected
Uninfected-1	—	232,450,106	212,391,460	—	—	—
Uninfected-2	—	235,048,212	203,760,783	—	—	—
Infected-1	37.5	234,378,088	199,871,662	10,219,315	4.86	13.0
Infected-2	26	226,078,422	198,436,507	7,322,556	3.56	13.7
Infected-3	21	233,750,850	205,747,441	7,241,973	3.40	16.2

Table 2.1: Samples used in this study, their infection rates and sequencing depth.

RNAs.

2.4.2 Changes in gene activity in primary T cells upon infection with HIV$_{89.6}$

Changes in host cell gene expression have been reported during HIV infection [68, 69, 111, 124, 129, 130] and differences in expression have been observed associated with the stage [140] and progression [151] of disease. Here we observed significant changes in gene expression (false discovery rate corrected $q < 0.01$) in 3,142 genes, 17.1% of expressed cellular genes (Additional file 1). The genes with most extreme increases, all $> 6\times$ fold higher, during HIV infection included IFI44L, RSAD2, HMOX1, MX1, USP18, IGJ, OAS1, CMPK2, DDX60, IFI44, IFI6, IFNG and CCL3. All of these have been reported to be involved in innate immunity [152] or are interferon inducible [153], highlighting a strong innate immune response in the cells studied. Genes with the largest decreases, all $> 3\times$ fold lower, were GNG4, GPA33, IL6R, CCR8, RORC, AFF2 and CCR2.

Many gene ontology categories were significantly enriched for differentially expressed genes (Additional file 2). Notably upregulated with infection were genes involved in apoptosis, immune responses and cytokine production (all $q < 10^{-4}$) and down-regulated were genes involved in viral gene expression, nonsense-mediated decay and translation elongation and termination (all $q < 10^{-19}$). These changes suggest that the cells responded to HIV infection with the induction of inflammatory, interferon regulated and apoptotic responses, patterns posited from several previous studies [70, 111, 124–130, 133, 154]. Several genes were activated that were characteristic

Cell type	HIV type	Differentially expressed genes (Up/Down)	Study
Primary $CD4^+$ T	$HIV_{89.6}$	3393 (1756/1637)	This study
Primary $CD4^+$ T	NL4-3 BAL-IRES-HSA	228 (182/46)	Imbeault et al. [129]
Lymph node biopsies	Acute infection	448 (383/65)	Li et al. [140]
SupT1	HIV_{LAI}	4997 (2666/2331)	Chang et al. [70]
SupT1	NL4-3Δenv-eGFP/VSV-G	579 (212/367)	Lefebvre et al. [111]

Table 2.2: Data from this study and four others used for meta-analysis of human gene expression changes during HIV infection

of other hematopoietic lineages, e.g. hemoglobin β, CD8, CD20 and CD117, while several $CD4^+$ T cell specific genes, e.g. CD4 and CD3, were downregulated, potentially consistent with de-differentiation of infected and bystander cells. We return to this point in the discussion.

2.4.3 Comparison of transcriptional profiles from $HIV_{89.6}$ infection of primary T cells to data on HIV infection in other cell types

We sought to identify the transcriptional responses that were most conserved upon HIV infection and so collected and analyzed data from four other studies of transcription in HIV-infected cells (Table 2.2). These included two studies of infection of the SupT1 cell line [70, 111], a study of primary $CD4^+$ T cells [129] and a study of lymphatic tissue in acutely viremic patients [140]. Genes were scored as increased or decreased in activity after infection, and the amount of agreement was compared among the different studies.

No gene was called as differentially expressed in all five studies. Eight genes were differentially expressed in the same direction in 4 out of 5 studies; AQP3 and EPHX2 were down-regulated with HIV infection and CD70, EGR1, FOS, ISG20, RGS16 and SAMD9L were up-regulated. A full listing is provided in Additional file 4. Several of the up-regulated genes are known to be interferon inducible, again emphasizing the role of innate immune pathways.

For each pair of studies, we compared whether they agreed on the identities of differentially expressed genes and whether they agreed on the direction of change (Figure

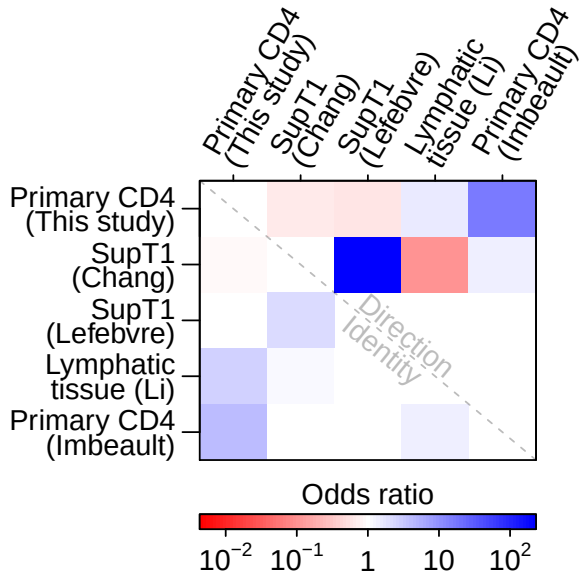


Figure 2.1: For each pair of studies, the association between up- and down-regulation calls was measured for genes identified by both studies as differentially expressed (above the diagonal). As another comparison, we also measured the agreement between studies for which genes were called differentially expressed regardless of direction (below the diagonal). The color scale shows the conservative (i.e. closest to 1) boundary of the confidence interval of the odds ratio with blue indicating a positive association and red a negative association between studies. For confidence intervals overlapping 1, the value was set to 1. Therefore all colored squares indicate significant associations.

2.1). The estimated alterations in gene activity showed notable differences in the responses to infection in primary cells versus the SupT1 cell line. The two SupT1 studies were significantly similar ($p < 10^{-15}$) to each other but were not significantly associated (Lefebvre et al. [111], $p = 0.2$) or were negatively associated (Chang et al. [70], $p = 10^{-7}$) with data from lymphatic tissue in acute HIV patients. The primary T cell study reported here was significantly associated with the second study in primary cells ($p < 10^{-15}$) and with a study of lymphatic tissue from patients acutely infected with HIV ($p = 0.003$). Our primary T cell data was negatively associated with the SupT1 studies (both $p < 10^{-3}$). This documents significant differences in responses to HIV infection between infected primary cells and SupT1 cells and suggests that results of infections in primary cells more closely align with actual acute HIV infections in patients. SupT1 cells might be expected to respond to infection differently that primary cells since they have several nonsynonymous mutations in innate immunity genes [155], have blocks in immune signaling pathways [156] and fail to activate many interferon stimulated genes during HIV infection [130].

2.4.4 Comparison of the HIV infected cell transcriptional profiles to additional experimental T cell profiles

To investigate the transcriptional changes in more depth, we compared the results of the five studies of HIV infection to transcriptional profiles comparing immune cell subsets available at the Molecular Signatures Database (MSigDB) [141]. The MSigDB reports genes that are increased or decreased in relative expression for each of 185 pairs of transcriptional profiles involving $CD4^+$ T cells. We compared the lists of affected genes in each pair to genes altered in activity by HIV infection. Those pairs of studies with the most significant associations with $HIV_{89.6}$ data are show in Figure 2.2A. For comparison, the associations with the four other HIV transcriptional profiling studies mentioned above are shown as well.

The most significant associations for our data showed gene expression in $HIV_{89.6}$-infected cells moving away from typical T cell expression patterns and towards patterns more similar to B cells, myeloid cells and bulk peripheral blood mononuclear cells (all Fisher's $p < 10^{-15}$) (Figure 2.2A). These changes were also seen, although to a lesser extent, in the Imbeault et al. [157] study which also used primary $CD4^+$ T cells.

For comparison, we also extracted those profiles most strongly associated with the transcriptional data on lymphatic tissue of HIV patients [140]. The profiles showed patterns similar to strongly stimulated T cells, autoimmune disease and to the Th1 T cell subset (all $p < 0.01$) (Figure 2.2B). Our data in primary $CD4^+$ T cells paralleled the changes seen in lymphatic tissue. These transcriptional changes again highlights the strong immune response generated by HIV infection in primary cells.

2.4.5 Intron retention

Cells respond to infection by shutting down macromolecular synthesis at multiple levels [158–162], so we investigated whether cells also showed perturbations in splicing efficiency after infection. As a probe, we created a database of cellular genomic regions

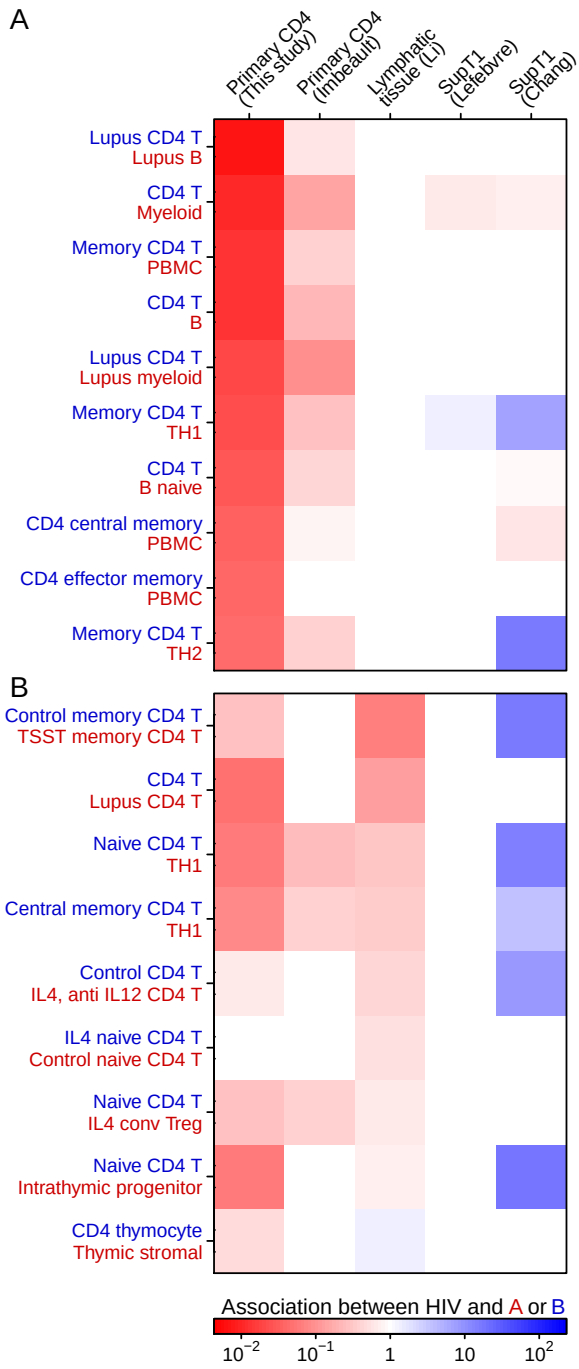


Figure 2.2: The MSigDB database was used to extract 185 sets of differentially expressed genes from pairs of transcriptional profiling studies of immune cell subsets involving $CD4^+$ T cells. For each pair of studies, we used Fisher's exact test to measure the association between up- and down-regulation calls for genes identified as differentially expressed in both our HIV study and the comparator immune subsets. A) The transcriptional profiles with strongest associations with changes observed in our study of $HIV_{89.6}$ infection of primary T cells. Blue indicates an positive association between changes seen in HIV infected cells and the first immune subset (text colored blue) while red indicates a positive association with the second immune subset (text colored red). The color scale shows the conservative (i.e. closest to 1) boundary of the confidence interval of the odds ratio. For confidence intervals overlapping 1, the value was set to 1. Therefore all colored squares indicate significant associations. B) As in A, but showing the transcriptional profiles most strongly associated with changes observed in lymph node biopsies from acutely infected patients [140].

annotated exclusively as exons or introns in all spliceforms in the UCSC gene database [163] and quantified expression in these regions in infected and uninfected cells. We found a significant increase in intronic sequences relative to exonic sequence (Wilcoxon $p < 10^{-15}$) (Figure 2.3A). This increase in intronic sequence was reproducible between replicates in our study (Kendall's τ=0.42, $p < 10^{-15}$) (Figure 2.3B). We reanalyzed RNA-Seq data from Chang et al. [70] and also documented intron retention which correlated with the changes seen in our data (Kendall's τ=0.12, $p < 10^{-15}$) (Figure 2.3C).

A possible artifactual explanation for enrichment of intronic sequences could involve greater DNA contamination in the infected cells samples. That is, if the relative amount of DNA differed between treatments, the amount of apparent intronic sequences could also differ due to sequencing of contaminating DNA. To examine whether DNA contamination was abundant in our samples, we compiled a collection of 27 large gene desert regions, defined here as 1) regions outside the centrosome and first and last cytoband, 2) containing less than 1% unknown sequence, 3) containing no genes annotated in UCSC genes [163], 4) containing no repeats annotated in the repeatMasker database [164] and 5) spanning more than 100kb. No reads were mapped to these 41Mb of gene deserts in any sample, arguing against explanations based on DNA contamination. Thus these data indicate that intron retention was increased in these cell populations upon HIV infection, revealing a previously undisclosed aspect of the host cell transcriptional response to infection.

Previous studies have reported changes in the expression and localization of splicing factors with HIV infection [132, 165, 166]. In our data, $\text{HIV}_{89.6}$ infection significantly altered the expression of genes involved in RNA splicing ($p = 2 \times 10^{-7}$) and nonsense-mediated decay ($p < 10^{-15}$). Genes related to nonsense-mediated decay genes showed a strong pattern of lowered RNA abundance, with 71 out of 118 annotated genes significantly lower in expression after infection. These patterns suggest potential mechanisms for the intron retention observed here.

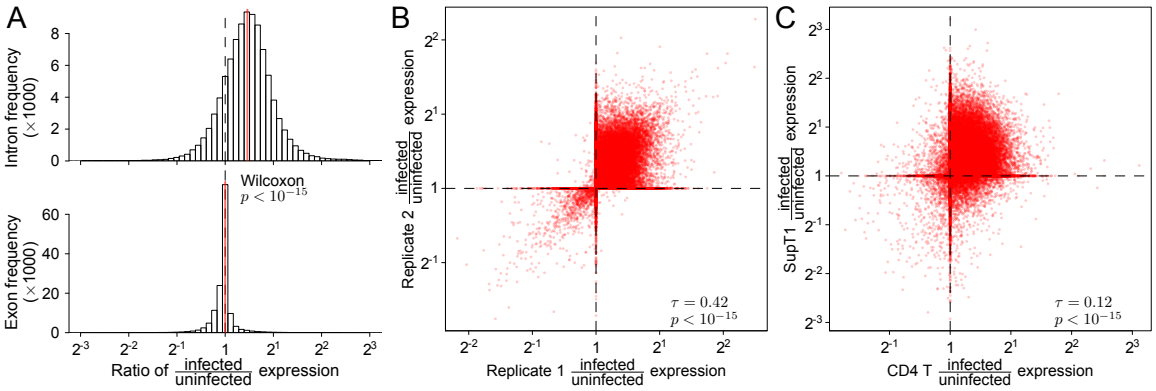


Figure 2.3: Expression of intronic and exonic regions was quantified as the proportion of reads mapping within the intron/exon out of the total reads mapping to the transcription units overlapping that intron/exon. A) Comparison of the ratios of expression between infected and uninfected replicates in exclusively intronic or exonic regions of transcription units. B) Reproducibility of intron retention between replicates. Each point quantifies the change in expression with HIV infection for a specific intronic region. The x-axis shows shows changes in gene activity accompanying infection for one set of replicates (Infected-1 and Infected-2 vs. Uninfected-1) and the y-axis shows the same data for different replicates (Infected-3 vs. Uninfected-2). C) Reproducibility of intron retention between studies. The plot is arranged as in B but with all data from our study combined on the x-axis and corresponding data from Chang et al. [70] on the y-axis.

2.4.6 Induction of transcription from HERVs and LINEs by $HIV_{89.6}$ infection

HIV infection has been reported to induce expression of certain HERVs, particularly HERV-K [167–169], and LINE and Alu transposable elements [170], providing candidate markers of infection and possible vaccine targets. Thus we analyzed our data in primary T cells infected with $HIV_{89.6}$ to investigate the expression of HERVs, LINEs and other repeated sequences. Figure 2.4A shows a comparison of the association between changes in expression with $HIV_{89.6}$ infection and the various genomic repeat types over varying levels of differential expression. At high levels of expression, ERV-9 (odds ratio at $4\times$ expression: 152, 95% CI:82.5–259) and its long terminal repeat LTR12C (odds ratio at $4\times$ expression: 144, 95% CI: 98.2–207) are the only repeats highly associated with upregulation during HIV infection. Looking at genomic repeats with any significant increase, the expression of many recently acquired genomic repeats, including L1HS, LTR5_Hs (a human specific LTR of HERV-K), AluYa5, AluYg6 and SVA_D and SVA_F, were associated with $HIV_{89.6}$ infection (Figure 2.4B).

We saw a relationship between the age of genomic repeats and its likelihood of being induced by $HIV_{89.6}$ infection. The most highly enriched repeats were associated with relatively recent hominid-specific repeat classes as annotated by the RepeatMasker database (repeat classes with $p < 10^{-50}$ odds ratio: 31.6, 95% CI: 8.88–112). In HERV-K (HML-2), the most recently active endogenous retrovirus in the human genome [144, 171, 172], we saw that integrations unique to the human genome [144] were more likely to be differentially expressed than older HERV-Ks (odds ratio: 5.38, 95% CI: 1.93–16.0).

Previous RNA-Seq studies of cellular expression during HIV infection in transformed cell lines did not report increases in HERV mRNA [70, 111]. To investigate this difference, we downloaded and analyzed the RNA-Seq data from Chang et al. [70], which quantified gene activity in transformed SupT1 cells infected with a lab-adapted strain of HIV. We found a much higher level of HERV expression in their data in

both HIV infected cells and uninfected controls than in primary cells (Figure 2.4C). We suspect that in SupT1 cells, as with many cancerous cells [173–177], the baseline expression of transposons and endogenous retroviruses is higher than in primary cells, masking further induction by HIV infection.

We observed heterogeneous expression among ERV-9/LTR12C sequences and so investigated the primary sequence determinants. We observed that ERV-9/LTR12C has three variants of differing length in the U3 region just upstream of the transcription start site (Figure 2.5A), an important region for transcription initiation [146]. The U3 region of LTR12C also contains multiple motifs for transcription factors NFY, GATA2 and MZF1 [149]. To clarify factors affecting expression levels, we counted the number of motifs matching these transcription factors, assigned each LTR12C to one of the length classes, counted the number of mutations away from the consensus for that length class and checked for integration in a transcription unit. We then carried out a regression analysis to test the effects of these variables on LTR12C differential expression. We found that $HIV_{89.6}$ induced transcription was more likely with the fewer mutations away from consensus, the number of locations matching the NFY transcription factor binding motif (CCAAT) and LTRs containing the short length variant of the 3′ U3 region. The presence of a MZF1 motif near the transcription start site decreased transcription (Figure 2.5B).

2.4.7 HIV mRNA synthesis and splicing

Over 24 million Illumina reads mapped to $HIV_{89.6}$, yielding an average coverage of over 240,000-fold. Reads mapping to $HIV_{89.6}$ comprised between 3.4–4.8% of mapped reads in the infected samples (Table 2.1). Assuming HIV-infected cells contain the same amount of mRNA as uninfected cells and adjusting for rates of infection ranging between 21–37.5% (Table 2.1), we estimate that HIV transcripts comprise between 13.0–16.2% of the total polyadenylated mRNA nucleotides in infected cells 48 hours after initial infection. This parallels previous estimates of around 10% [178] at 48 hours

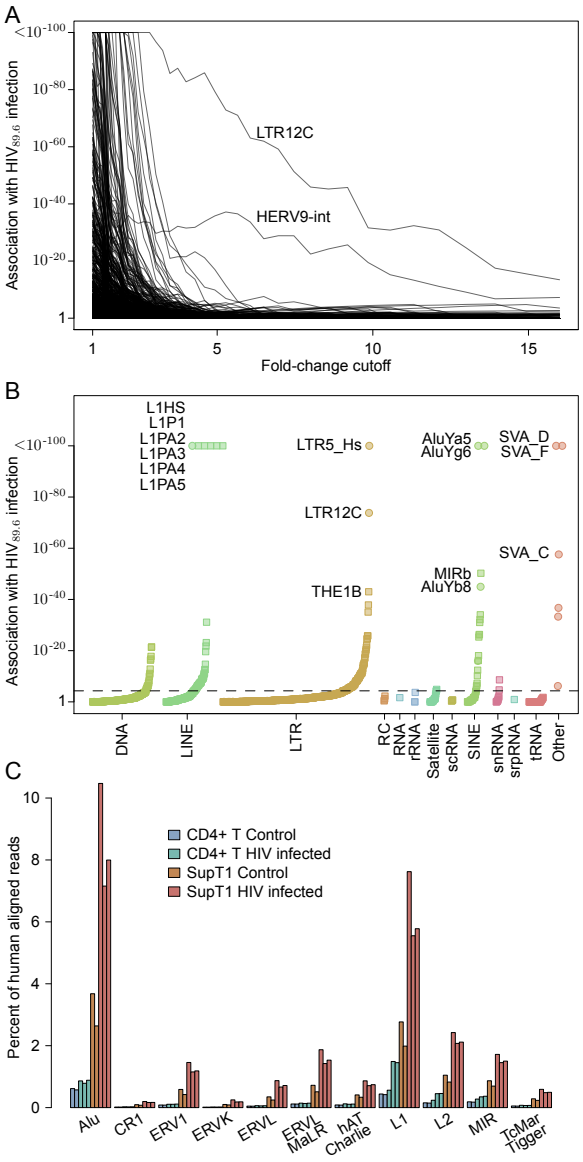


Figure 2.4: A) The association of repeat regions differentially expressed after $HIV_{89.6}$ infection of primary T cells observed for varying thresholds of differential expression. The threshold used to call a gene differentially expressed based on the Bayesian posterior median was varied and Fisher's exact test was used to assess whether any genomic repeats had a significant association with this differential expression. Note that only ERV-9 (annotated as HERV9-int in the RepeatMasker database) and it's corresponding long terminal repeat LTR12C were significantly associated with large changes in expression. B) Enrichment of repeat categories in regions differentially expressed (Bayesian 95% credible interval > 1) between HIV-infected and control $CD4^+$ T cells. The repeated sequences are ordered on the x-axis by the extent of induction within each class, the y-axis shows the p-value for upregulation after infection. The dashed line indicates a Bonferroni corrected p value of 0.05. (C) The proportion of human mapped reads that align within classes of genomic repeats for data from primary $CD4^+$ T cells from this study and SupT1 cells from Chang et al. [70]. A single read mapping multiple times to a given category was only counted once.

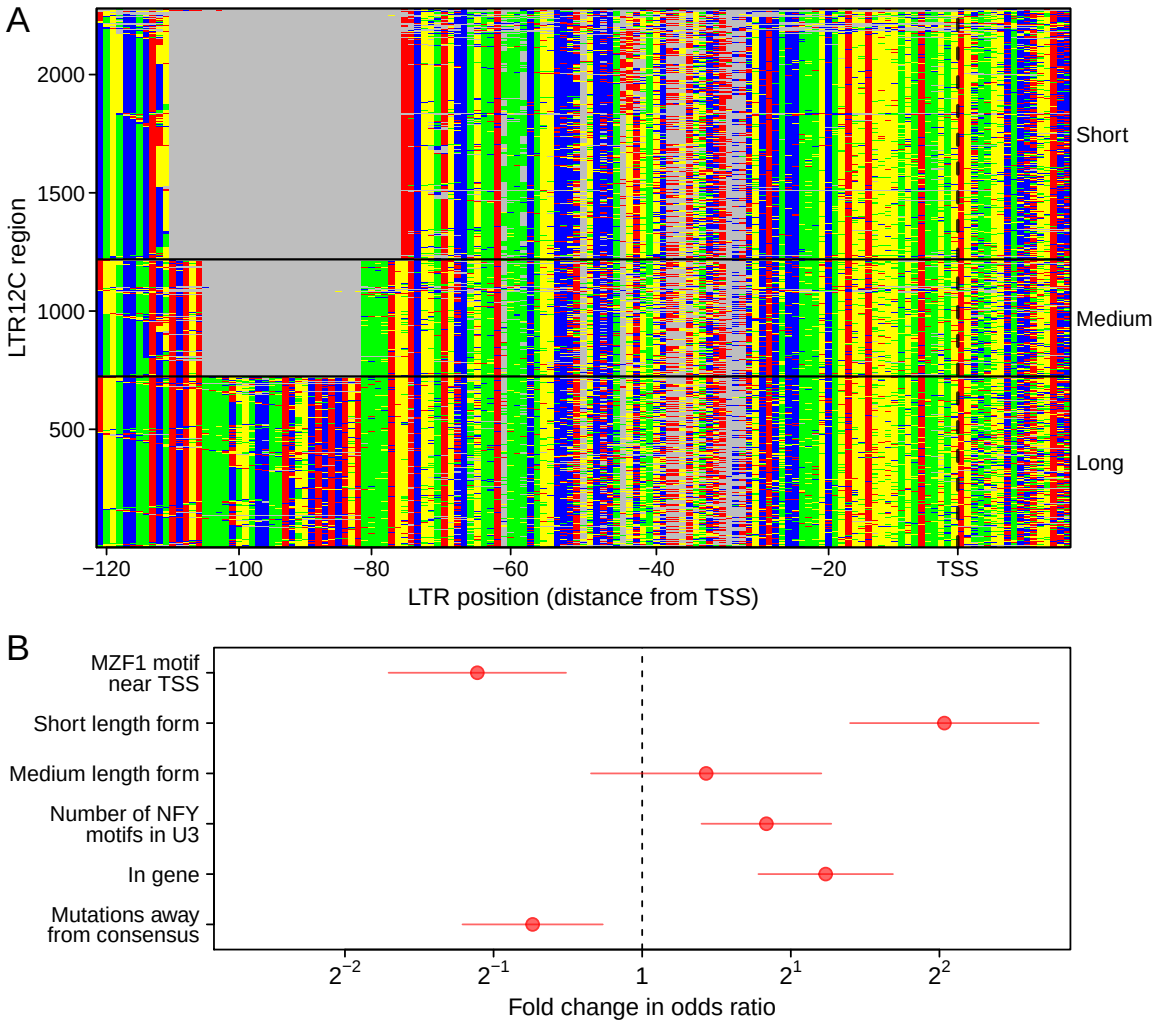


Figure 2.5: A) An alignment of the 3′ end of the U3 region of repeats annotated as ERV-9 LTR12C. Each row is a LTR sequence and each column a base in that sequence colored by nucleotide identity. Three distinct classes are visible with a short, medium and long form. Mutations away from the consensus can also be seen. B) The coefficients (points) and ± 1.96 standard errors (horizontal lines) of a logistic regression comparing differential expression of LTR12C to the presence of MZF1 and NFY motifs, short/medium/long length alternate forms of the U3-R region, mutations away from the consensus for each length form and integration inside a transcription unit. The coefficient shown for mutations away from consensus is for a 10 mutation difference and the coefficient shown for NFY motifs is for a change of 5 additional motifs. All other coefficients are for binary values.

postinfection, 38% at 24 hours [70] or 30% after 72 hours [124].

Over 47,257 single reads spanned previously reported HIV splice junctions, allowing a quantitative assessment of donor and acceptor utilization (Figure 2.6A). As expected from previous studies [57, 114], the most abundant junctions were D1-A5 and D4-A7. We confirmed the use of unusual splice acceptors A8c and A5a, previously reported in $\text{HIV}_{89.6}$ [114]. In our data, we also see a higher abundance of D1-A1 and D1-A2 splice junctions than might be expected [57, 114], although previous studies reported proportional abundance within size classes, making comparisons between size classes uncertain.

A 3′ bias is apparent in our sequencing data (Additional file 5). This could be due to the poly-A capture step of the protocol where any break in the RNA would result in distal 5′ sequences being lost [179]. We used sequence reads from the large unspliced HIV intron 1 to measure this bias using a regression of the $\log$ of the number of fragments with a 5′-most end starting at a given position against the distance of that position from the viral polyadenylation site, yielding an estimated probability of breakage of 0.021% per base (Additional file 5). Given this rate of termination, there is only a 14% chance of reaching the 5′ end of the 9171nt unspliced HIV genome ($(1 - 0.00021)^{9171}$).

Ocwieja et al. [114] determined the relative abundance of $\text{HIV}_{89.6}$ of similarly sized transcripts using PacBio single molecule sequencing, but were not able to estimate the relative abundance of all transcripts due to a sequencing bias favoring shorter transcripts. For this reason, relative abundances could only be specified within message size classes (i.e. the 4kb, 2kb and unexpectedly a 1kb size class as well) and the overall quantitative abundances were unknown. The RNA-Seq data reported here are unable to determine complete transcript abundance because the short read length does not allow reconstruction of multiply spliced messages but do permit estimation of size class abundances after correcting for 3′ bias (Additional file 5). Thus the PacBio data reported by Ocwieja et al. [114] and the Illumina data reported here can be combined together to

determine complete relative abundance of all $HIV_{89.6}$ transcripts (Figure 2.6B).

The most abundant HIV mRNAs were the unspliced HIV genome (37.6%), a transcript encoding Nef (D1-A5-D4-A7: 15.5%), two 1kb size class transcripts (D1-A5-D4-A8c: 10.6%, D1-A8c: 4.9%) and two Rev-encoding transcripts (D1-A4c-D4-A7: 4.2%,D1-A4b-D4-A7: 3.1%). The function of this large amount of 1 kb transcript is unknown. These two 1 kb transcripts do not appear to encode significant open reading frames although other 1kb transcripts can encode a Rev-Nef fusion [114].

Using these abundances, we can estimate the number of $HIV_{89.6}$ genomes in these primary T cells 48 hours after infection. To determine the proportion of the mRNA nucleotides from viral transcripts, we multiplied the estimated abundances by their transcript lengths. Unspliced genome transcripts appear to form 79% of the mRNA nucleotides from $HIV_{89.6}$ transcripts. Assuming T cells contain at least 0.1pg of mRNA then an infected cell should contain at least 0.011pg of unspliced HIV transcript ($0.1\text{pg} \times 0.14\frac{\text{HIV mRNA nt}}{\text{cell mRNA nt}} \times 0.79\frac{\text{unspliced mRNA nt}}{\text{HIV mRNA nt}}$) or, assuming 9171 bases of RNA weigh about 5×10^{-6} pg, at least 2200 HIV genomes at 48 hour post infection. This estimate roughly agrees with previous estimates of HIV production per cell [178, 180, 181].

2.4.8 Human-HIV chimeric reads

The suggestion that HIV integration may disrupt cellular cancer-associated genes and thereby promote cell proliferation [182–185] has focused attention on the range of novel message types formed when HIV integrates within transcription units [186–190]. Chimeric reads containing HIV and cellular sequence are also of clinical interest due to the potential of lentiviral vectors to trigger oncogenesis in gene therapy patients through insertional mutagenesis [191–194].

In our data, 80,045 reads contained sequences matching to both HIV and human genomic DNA, but a considerable complication arises because chimeras can be formed artifactually during the preparation of libraries for sequence analysis [195–202]. Many

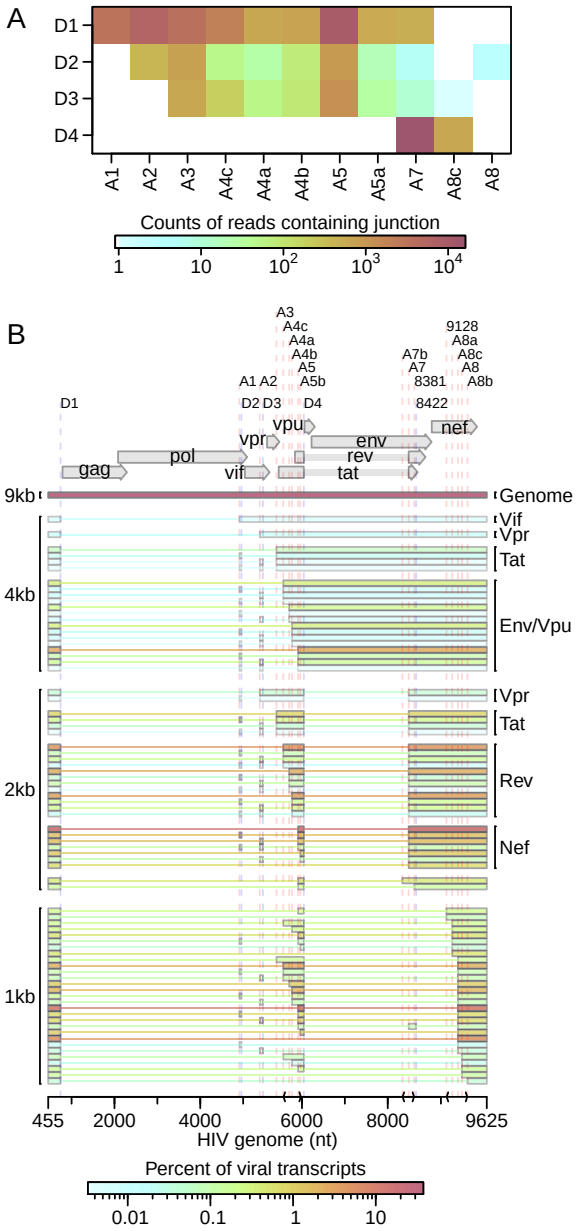


Figure 2.6: A) Junctions between HIV splice donors and acceptors observed in the RNA-Seq data. Acceptors are shown as the columns and donors as the rows with the coloring indicating the frequency of each pairing. B) The relative abundance of all $HIV_{89.6}$ transcripts as determined by a combination of PacBio sequencing [114] and Illumina sequencing. Message structures were generated by targeted long read single molecule sequencing, which allowed association of multiple splice junctions in single sequence reads. The Illumina short read sequencing allowed normalization of message abundances between size classes. The inferred HIV message population is shown colored by relative abundance.

of the chimeric sequences in our data contained junctions between the HIV and human sequence where the ends of the human and HIV sequence were similar and potentially complementary (Figure 2.7A). This raises the concern that some of these chimeras could be products of in vitro recombinations during the reverse transcription, amplification and sequencing processes. Template switching between sequences with shared similarity is a well established property of retroviral reverse transcriptase enzymes used in RNA-Seq library preparation [203–205]. Priming off incomplete transcripts during DNA synthesis is another potential source of chimeric transcripts [195, 196, 206, 207]. Failing to account for chimeras can hinder interpretation of deep sequencing data [197–202].

Also consistent with artifactual chimera formation, 7,354 reads (9.2% of chimeric messages) contained HIV sequences joined to human mitochondrial sequences, yet HIV proviruses have not previously been found integrated in mitochondrial DNA [187]. To probe this further, we used ligation-mediated PCR to recover integration site junctions from the same infected cell populations analyzed by RNA seq, yielding 147,281 unique integration sites (Figure 2.7B) [135]. No integrations in mitochondrial DNA were detected. We conclude that chimeric HIV-mitochondrial sequence reads in the RNA-seq data represent artifacts of library construction and so used these chimeras as an assay to evaluate subsequent data filtering steps. We reasoned that reads without sequence similarity at junctions between human and HIV mapping were less likely to be artifacts caused by template switching. Filtering to only reads where no overlap and no unknown intervening sequence was present between human and HIV portions left 2181 junctions and reduced the proportion of reads containing mitochondrial DNA to 2.4%. Of the remaining HIV-human chimeric reads, the HIV portion of 605 sequences bordered the 3′ or 5′ end of HIV or an HIV splice donor or acceptor. Filtering to these more likely authentic junctions left only 2 (0.3%) chimeric reads containing mitochondrial sequence. This decrease in likely mitochondrial artifacts suggests that the filtering was effective. The high rate of mitochondrial chimeras in the unfiltered sequences raises the concern

that artifacts may easily distort results in studies using similar amplification and sequencing techniques.

Chimeric messages composed of HIV and cellular RNA sequences can be formed by cellular gene transcription reading into the integrated provirus, by HIV transcription reading out through the viral polyadenylation site or by splicing between human and viral splice sites. In our filtered data, the predominant forms appear to be derived from reading through the HIV polyadenylation signal into the surrounding DNA (78%), splicing out of the viral D4 splice donor to join to human slice acceptors (17%) and reading into the HIV 5′ LTR from human sequence (4.0%) (Figure 2.7C). No splice site other than D4 had more than two chimeric reads observed.

The filtered chimeric reads had many traits consistent with biological chimera formation. The reads containing HIV D4 joined to human sequences had the characteristics expected of splicing—72.1% of the chimeric junctions mapped to known human acceptors and 96.1% mapped to a location immediately preceded by the AG consensus of human mRNA acceptors. The reads containing the 5′ or 3′ LTR border were almost exclusively (93%) found in transcription units, with odds of being in a gene 2.3-fold (95% CI: 1.6–$3.2\times$) higher than integration sites from the same sample. The 5′ or 3′ chimeras were also more likely to be located in an exon than integration sites even after excluding any integration or chimera not located in a transcription unit (odds ratio: $2.1\times$, 95% CI: 1.6–$2.6\times$).

We next compared whether the human and viral segments of chimeric reads agreed or disagreed in orientation (i.e. strand transcribed) for reads with the human portion mapped within annotated transcription units. The sequencing technique used here does not preserve strand information, but we can check whether the strand of a sequence read agrees or disagrees with the annotated gene strand and compare this to the observed strand of the HIV portion of the read. We found a strong association between the orientation of the human and HIV portions of chimeric reads within 3′ and 5′

chimeras (odds ratio: $6.2\times$, 95% CI: 3.9–$10.2\times$). This highly significant enrichment of HIV and human genes in the same orientation (Fisher's exact test $p < 10^{-15}$) might indicate that antisense HIV RNA is rapidly degraded by a response to double-stranded RNA or that polymerases oriented in opposing directions interfere with one another during elongation. Chimeras involving HIV splice donor D4 were even more highly enriched for matching orientations (odds ratio: $52.5\times$, 95% CI: 12.1–$307\times$) suggesting that pairing with human splice acceptors may add an additional constraint on the orientation of D4 chimeric reads.

Based on these data, we can propose a lower bound on the relative abundance of chimeras. If we assume that our filtering removed nearly all artifacts so that we have few false positives, then our estimate should be lower than the true proportion of chimeras. In our data, only $\frac{604}{12{,}689{,}879} = 0.0048\%$ of reads containing sequence mapping to HIV also contained identifiable chimeric junctions. However, this is an underestimate because in an HIV-derived mRNA, any fragment of the sequence will be mappable to HIV, while for a chimeric sequence only a read spanning the HIV-human junction will allow identification of a chimera. If we assume that 25 bases of sequence are necessary to map to human or HIV sequence, then, with the 100-bp reads used here, only read fragments starting between 75- and 25-bp downstream of the chimeric junction will be identifiable. If we assume the average chimeric mRNA sequences is at least 2kb long, then a read from a chimeric sequence has at most a $\frac{50}{2000} = 2.5\%$ chance of containing a mappable junction. Thus, a lower bound for the proportion of HIV mRNA that also contain human-derived sequences is 0.2% ($\frac{0.0048\%}{2.5\%}$). Looking only at splicing from HIV donor D4, we saw 16,843 reads containing a junction from D4 to an HIV acceptor and 104 reads from D4 to human sequence. Thus, in our data, 0.6% of D4 splice products form junctions with human acceptors instead of HIV acceptors.

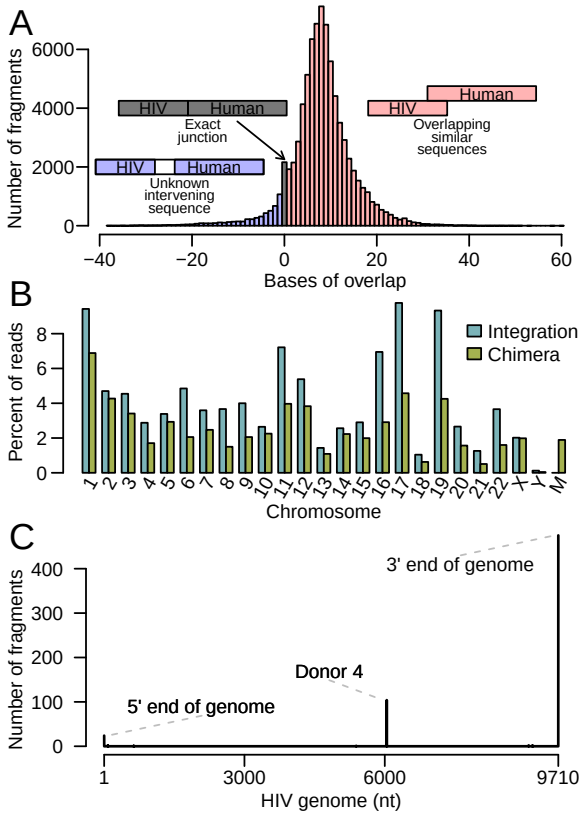


Figure 2.7: A) The length of overlapping sequence (regions of complementarity potentially favoring chimera formation) matching both human and HIV at inferred chimeric junctions. The x-axis shows the length of the overlap and the y-axis shows the frequency of chimeric junctions with the indicated extent of overlap. B) Chromosomal distribution of uniquely mapping HIV integration sites from the same infections of primary T cells and comparison to uniquely mapping human sequences in chimeric reads observed in RNA-Seq. Note that the mitochondrial genome, denoted as M, has no authentic integration sites but does have extensive matches to chimeric junctions found in the RNA-Seq data. C) Counts of the location in the HIV genome of the HIV-human junctions in filtered chimeric reads.

2.5 Discussion

Here we used RNA-Seq to analyze mRNA accumulation and splicing in primary T cells infected with the low passage isolate $HIV_{89.6}$. We did not carry out dense time series analysis, compare different human cell donors or compare different perturbations of the infections—instead, we focused on generating a dense data set at a single time point. We analyzed replicate infected cell and control samples to allow discrimination of within-condition versus between-condition variation and assessed differences using a series of bioinformatic approaches. Many previous studies have used microarray technology or RNA-Seq to study gene activity in HIV-infected cells [68, 70, 111, 124–131, 133], usually analyzing infections of transformed cell lines or laboratory adapted strains of HIV-1. Here we present what is to our knowledge the deepest RNA-Seq data set reported for infection in primary T cells using a low passage HIV isolate ($HIV_{89.6}$). This data set was paired with a set of 147,281 unique integration site sequences extracted from the

same infections, which were critical to our ability to quality control chimeric reads. An advantage of studies using cell lines and laboratory adapted strains is that often a high percent of cell infection can be achieved, whereas in this study we achieved only around $\sim 30\%$ infection. However, we report distinctive features of the transcriptional response not seen in studies of HIV infections in cell lines. Novel in this study are 1) identification of intron retention as a consequence of HIV infection, 2) the finding of activation of ERV-9/LTR12C after HIV infection, 3) generation of a quantitative account of the structures and abundances of over 70 $\text{HIV}_{89.6}$ messages and 4) clarification of the predominant types of HIV-host transcriptional chimeras. These findings are discussed below.

Broad changes in host cell mRNA abundances were evident after infection, with over 17% of expressed genes changing significantly in activity. Changes included expected response to viral infection, apoptosis and T cell activation. Although it is not possible here to separate the response of infected and bystander cells, this study highlights the drastic changes in cellular expression caused by HIV-1 infection. In a meta-analysis including four previously published studies, no gene was detected as differentially expressed in all five studies and only a handful of genes appeared in four out of five studies. Further analysis showed that expression changes appear to be cell type specific, raising concerns that studies using cell lines may not fully reflect host cell responses in *in vivo* infections.

Unexpectedly, intronic sequences were more common in the RNA-Seq data from cells after $\text{HIV}_{89.6}$ infection than in mock infected cells. The mechanism is unclear. It is possible that the splicing machinery is reduced in activity after 48 hours of infection, perhaps as a part of the antiviral response of infected and bystander cells. HIV infection does appear to alter expression and localization of some splicing factors [132, 166]. In addition, we saw a large reduction in the abundance of mRNA from nonsense-mediated decay related genes, perhaps indicating that RNA surveillance is loosened

thus allowing more unspliced or aberrantly spliced transcripts. Alternatively, fully spliced mRNAs might be more rapidly degraded after infection, possibly by interferon-mediated induction of RNaseL [208]. A speculative possibility is that $HIV_{89.6}$ encodes a factor that alters cellular splicing or promotes mRNA degradation to optimize splicing and translation of viral messages.

Infection resulted in increased expression of specific cellular repeated sequences. HERVs, in particular HERV-K, have previously been observed to show increased RNA accumulation with HIV infection [167–169, 209] and possibly represent vaccine targets because of their production of distinctive proteins [173, 209–213]. Here, though we saw modest increases in HERV-K expression, ERV-9 had the greatest change in expression (33 LTR12C and 14 ERV-9 annotated regions with greater than $4\times$ change in expression). Previous RNA-Seq studies of HIV infection in cell lines did not report increases in HERV expression [70, 111] but this difference is likely due to a much higher baseline expression of HERVs in transformed cell lines. We also observed increases in LINE and Alu element transcription, as has been reported previously [170], and expression changes in ERV-9/LTR12C expression associated with transcription factor motifs and U3 variants.

Many of the repeated sequence elements that were induced by $HIV_{89.6}$ infection are relatively recently integrated in the human genome. The reason for this pattern is unclear. It may be that older elements have accumulated more mutations, resulting in an inactivation of transcriptional signals. Alternatively, perhaps the elements that are induced have been recruited for transcriptional control of cellular functions, so that their transcriptional activity is preserved evolutionarily [148, 214, 215].

Comparison of results of sequencing $HIV_{89.6}$ messages using long-read single molecule sequencing (Pacific Biosciences) and dense short read sequencing (Illumina data reported here) allowed a full quantitative accounting of more than 70 $HIV_{89.6}$ splice forms. The full length unspliced HIV RNA comprised 37.6% of all messages, corre-

sponding to about 2000 genomes per cell. Notably abundant messages included those encoding Nef (D1-A5-D4-A7: 15.5%) and two Rev-encoding transcripts (D1-A4c-D4-A7: 4.2%,D1-A4b-D4-A7: 3.1%). The full set of messages is summarized in Figure 2.6B. Our previous analysis revealed an unusually prominent 1 kb size class. $HIV_{89.6}$ encodes a rare splice acceptor (A8c) within Nef responsible for formation of the short messages. Our data indicated that two members of the 1-kb size class, D1-A5-D4-A8c and D1-A8c, accounted for 10.6% and 4.9% of all messages. The 1kb size class as a whole accounted for fully 20% of messages. Most HIV/SIV variants appear to encode an acceptor near this position, suggesting a potential unknown function for these short spliced forms [114, 216, 217].

After filtering, we detected a sizeable number of apparently authentic chimeras containing both HIV and cellular sequences, allowing comparison to examples of host-cell modification by integration. Mechanisms of insertional activation have been studied intensively in animal models of transformation and in adverse events in human gene therapy. One of the most common mechanisms involves insertion of a retroviral enhancer near a cellular promoter, so that the rate of initiation is increased and normal cellular messages are increased in abundance. However, another common mechanism involves formation of chimeric messages involving both cellular and viral/vector sequences. In HIV infection, examples of insertion in the Bach2 and MKL2 genes have been associated with long term persistence of particular cell clones [182–185]. In these cells, proviruses were integrated within the cellular transcription unit, and the transcriptional direction of the integrated provirus was the same as that of Bach2 or MKL2. This would allow formation of a fusion of the 5′ HIV sequences with 3′ Bach2 sequences, potentially involving the most common events seen here (either 3′ read out or splicing from HIV D4 to a cellular exon). However, a closely studied example of clonal expansion in a successful lentiviral vector gene therapy for beta-thalassemia was associated with expansion of a cell clone harboring an integrated vector within the transcription unit of HMGA2. In this case the message spliced into the vector and terminated, removing a

negative regulatory sequence normally present in the 3′ end HMGA2 message [191]. A targeted study in vitro of chimeric message formation by lentiviral vectors showed examples of multiple types of read-in and -out and splice-in and -out [193], which may have been more frequent and more varied than for $HIV_{89.6}$ proviruses studied here. The lack of splicing or reading into HIV in this study may be a reflection of the high rate HIV transcription in these infected cells–because HIV was so highly expressed, there would be more opportunities for polymerase to splice out of or read through the HIV genome than to read or splice in. The vast majority of HIV proviruses in expanded clones in well-suppressed patients now appear to be defective [185]—going forward, it will be of interest to investigate whether these HIV proviruses are damaged in ways that promote formation of chimeric transcripts.

Lastly, we note that several features of the transcriptional response to $HIV_{89.6}$ infection were suggestive of de-differentiation away from T cell specific expression patterns. The increase in expression of cellular HERVs and LINEs is characteristic of cells in early development. Specific HERVs and transposons, including ERV-9/LTR12C and HERV-K, have been implicated in regulating gene activity early in development [148, 214, 218–221]. Several genes related to other hematopoietic cell types showed elevated RNA abundance after $HIV_{89.6}$ infection. These data are of interest given the finding that patients undergoing long term ART can contain long lived T cell clones that may contribute to the latent reservoir [185, 222–225]. Possibly the transcriptional responses seen in infected primary T cells here are reflective of processes leading to formation of the long-lived latently-infected cells with stem-like properties.

2.6 Conclusions

Infections of primary T cells with a low passage HIV isolate show several distinctive features compared with previously published data using T cell lines and/or lab-adapted HIV strains. We found strong changes in expression in genes related to immune response and apoptosis similar to studies of HIV infection in patient samples and primary

cells but different from studies performed in SupT1 cell lines. Notable changes after infection included intron retention and activation of recently integrated retrotransposons and endogenous retroviruses, in particular LTR12C/ERV-9. We also present complete absolute estimation of over 70 messages from $HIV_{89.6}$ and specify the major virus-host chimeras as read out from the 3′ end of the provirus and splicing from viral splice donor 4 to cellular acceptors.

2.7 Availability of supporting data

RNA-Seq reads from this study are available at the Sequence Read Archive under accession number SRP055981. The integration site data is available at the Sequence Read Archive under accession number SRP057555.

2.8 Author's contributions

KEO performed the infections and sequencing. SS-M analyzed the data. SS-M, KEO and FDB planned the overall study, and SS-M and FDB wrote the paper. All authors read and approved the final manuscript.

2.9 Acknowledgements

We would like to thank the University of Pennsylvania Center for AIDS Research (P30 AI045008) for preparation of viral stocks and isolation of primary $CD4^+$ T cells; Ronald G. Collman and members of the Bushman laboratory for reagents, helpful discussion and technical expertise. This work was funded by NIH grant R01 AI052845, the HIV Immune Networks Team (HINT) consortium P01 AI090935 and NRSA computational genomics training grant T32 HG000046.

2.10 Additional Files

2.10.1 Additional file 1 — Analysis of genes differentially expressed during $HIV_{89.6}$ infection of primary $CD4^+$ T cells

Output from CuffLinks analysis of the RNA-Seq data organized in a csv file with columns UCSC gene ID, gene symbol, status of test, FPKM in uninfected and infected samples, the $\log_2$ fold change, test statistic and false discovery rate adjusted p-value.

2.10.2 Additional file 2 — Analysis of Gene Ontology categories associated with differential expression during $HIV_{89.6}$ infection of primary $CD4^+$ T cells

Counts of differentially expressed genes for each Gene Ontology category. Columns are the name of the category, the numbers of genes differentially up- and downregulated or not significantly changed and odds ratios and p-values from Fisher's exact tests.

2.10.3 Additional file 4 — Genes called as up- or downregulated by studies of expression during HIV infection

Genes called as differentially expressed in the five studies analyzed in the meta-analysis of differential expression with HIV infection. Columns are the study, the gene name(s) and whether the differential expression was up or down.

2.10.4 Additional file 5 — Estimating relative abundance of $HIV_{89.6}$ message size classes using RNA-Seq data

A) RNA-Seq coverage of the $HIV_{89.6}$ genome for the replicates in this study. Each replicate is indicated by a different color. The HIV genome is shown on the x-axis and the number of reads that aligned to each position is shown on the y-axis. Black line indicates the 0.021% coverage decrease per base distance from the $3'$ end of the mRNA estimated from a least squares fit on the read counts in the first intron. B)

Diagram of the segments of the $HIV_{89.6}$ RNA present in each of 9kb, 4kb, 2kb and 1kb size class. C) The proportion of reads mapped to each of the segments of the $HIV_{89.6}$ genome shown in B adjusted by the length of the segment. Each replicate is shown by a different color. D) Corrected representation of RNA segments from the different size classes. Because cDNA synthesis was primed from the polyA tail, more 3′ sequences are recovered preferentially. Using the bias estimate from A, we adjusted each genome segment by the inverse of the bias predicted based on its distance from the 3′ end of the mRNA. Corrected proportions for the indicated RNA segments are shown colored by replicate. E) The proportion of each size class was inferred using the estimates in D by calculating the difference between segments. Replicates are indicated by color.

CHAPTER 3 : HIV latency and integration site placement in five cell-based models

3.1 Abstract

Background: HIV infection can be treated effectively with antiretroviral agents, but the persistence of a latent reservoir of integrated proviruses prevents eradication of HIV from infected individuals. The chromosomal environment of integrated proviruses has been proposed to influence HIV latency, but the determinants of transcriptional repression have not been fully clarified, and it is unclear whether the same molecular mechanisms drive latency in different cell culture models.

Results: Here we compare data from five different *in vitro* models of latency based on primary human T cells or a T cell line. Cells were infected in vitro and separated into fractions containing proviruses that were either expressed or silent/inducible, and integration site populations sequenced from each. We compared the locations of 6,252 expressed proviruses to those of 6,184 silent/inducible proviruses with respect to 140 forms of genomic annotation, many analyzed over chromosomal intervals of multiple lengths. A regularized logistic regression model linking proviral expression status to genomic features revealed no predictors of latency that performed better than chance, though several genomic features were significantly associated with proviral expression in individual models. Proviruses in the same chromosomal region did tend to share the same expressed or silent/inducible status if they were from the same cell culture model, but not if they were from different models.

Conclusions: The silent/inducible phenotype appears to be associated with chromosomal position, but the molecular basis is not fully clarified and may differ among *in vitro* models of latency.

3.2 Background

Highly active antiretroviral therapy (HAART) can suppress HIV-1 replication in infected patients, but the ability of HIV to persist as an inducible reservoir of latent proviruses [226–228] obstructs eradication of the virus and functional cure [229]. These latent proviruses are long lived [230, 231] and relatively invisible to the immune system [227, 232]. The potential for even a single virus to restart infection despite successful antiviral therapy means that it may be necessary to eliminate all latent proviruses to eradicate HIV from an infected person.

After integration, a positive feedback loop of Tat transactivation appears to partition proviral gene activity into either of two stable states [233–235]—abundant Tat driving high proviral expression or little Tat leading to quiescent latency. Similar to the positional effect variegation observed in fruit fly chromosomal rearrangements [236, 237], studies on cell clones with single integrations show that differing integration sites can have large differences in proviral expression [238–240]. These data suggest that integration site location, along with the cellular environment [240–243], influences the balance between latency and proviral expression.

Associations between latency and genomic features have also been reported in collections of integration sites from cell culture models although the consistency of these effects across model systems and their relationships to latency in patients remains uncertain. Lewinski et al. [244] reported that proviruses integrated in gene deserts, alphoid repeats and highly expressed genes are more likely to have low expression. Shan et al. [245] reported an association between latency and integration in the same transcriptional orientation as host genes. Pace et al. [246] found that silent and expressed provirus integration sites differed in the abundance and expression levels of nearby genes, GC content, CpG islands and alphoid repeats. In model systems with defined integration sites, Lenasi et al. [247] reported decreased and Han et al. [248] reported increased viral transcription when the provirus is downstream of a highly

expressed host gene.

Cell-based models of latency are important for many aspects of HIV research, including screening small molecules that can reverse latency and potentially allow eradication [249, 250]. Location-driven differences in expression are preserved even after demethylation and histone deacetylase treatment [238], which suggests that integration location has the potential to confound "shock and kill" anti-latency treatments [251, 252]. A greater understanding of the effects of integration site location on latency could thus affect antiretroviral development.

To search for features of integration site associated with latency, we generated a set of inducible and expressed integration sites using a primary central memory $CD4^+$ T cell model of latency [253, 254], collected four previously reported integration site datasets and modeled the effects of genomic features near the integration site on the expression status of these proviruses. Although some genomic features associated with latency in individual models, no feature was consistently associated with proviral expression across all five cell culture models. However, closely neighboring proviruses within the same cellular model shared the same latency status much more often than expected by chance suggesting that chromosomal position of integration affects latency but that the mechanism remains unclear or differs between cell culture models. Thus these data help inform the design of experiments in HIV eradication research.

3.3 Methods

3.3.1 Integration sites

Naive $CD4^+$ T cells were purified by negative selection from peripheral blood mononuclear cells. The cells were activated with anti-CD3 and anti-CD28 (+TGF-beta, anti-IL-12, and anti-IL-4) to generate "non-polarized" cells (the in vitro equivalent of central memory T cells). Five days after isolation, cells were infected with an NL4-3-based virus with GFP in place of Nef and the LAI envelope (X4) provided in trans at a concentration

of 500 ng of p24 as measured by ELISA per million cells. Based on previous experience with this model, this amount of p24 should produce an MOI of approximately 0.15. Cells were cultured in the presence of IL-2. Two days post-infection, cells were sorted for GFP+; this active population expresses GFP even when treated with flavopiridol, although for this study they were not treated. The inducible population was the set of GFP negative cells from the initial sort that, 9 days post-infection, were activated with anti-CD3 and anti-CD28 and sorted for GFP production.

Genomic DNA from the inducible and expressed populations was digested with MseI, ligated to an adapter, and amplified by ligation-mediated PCR essentially as in Wu et al. [255] and Mitchell et al. [256] except that the nested PCR primers included sequence for the Ion Torrent P1 adapter and adapter A sequence with a 5 base barcode sequence specific to the inducible or expressed conditions. Amplicons were sequenced using an Ion Torrent Personal Genome Machine (PGM) according to manufacturer's instructions using an Ion 316 chip and the Ion PGM 200 Sequencing kit (Life Technologies). The sequence reads were sorted into samples by barcode. All reads were required to match the expected $5'$ sequence with a Levenshtein edit distance less than 3 from the expected barcode, $5'$ primer and HIV long terminal repeat (LTR). The $5'$ primer and HIV sequence, along with the $3'$ primer if present, were trimmed from the read. Sequences with less than 24 bases remaining or containing any eight base window with an average quality less than 15 were discarded. Duplicate reads and reads forming an exact substring of a longer read were removed.

3.3.2 Analysis

All statistical analysis was performed in R 2.15.2 [139]. The analyses are described in a reproducible report (Appendix A.1). The annotated integration site data necessary to perform the analyses and the compilable code to generate this reproducible report are provided as supplemental information [189]. The new Central Memory $CD4^+$ data set was analyzed as in Berry et al. [257]. The integration patterns appeared similar to

previously reported HIV integration site datasets [258].

3.3.3 Previously published data

We collected integration sites from three previously reported studies (Table 3.1), for a total of four expressed versus silent/inducible pairs of samples. These studies used primary $CD4^+$ T cells or Jurkat cells infected with HIV or HIV-derived constructs as cell culture models of latency. Flow cytometry allowed cells expressing viral encoded proteins to be sorted from non-expressing cells. In two of the studies, these non-expressing populations were stimulated to ensure that the provirus could be aroused from latency. Specific differences in protocol between the study sets are summarized below.

Jurkat. Lewinski et al. [244] infected Jurkat cells with a VSV-G pseudotyped, GFP-expressing pEV731 HIV construct (LTR-Tat-IRES-GFP) [238] at an MOI of 0.1. The cells were sorted into GFP+ and GFP- two to four days after infection. GFP+ cells were sorted again two weeks after infection and cells that were again GFP+ were collected for integration site sequencing. GFP- cells were sorted for GFP negativity twice more then stimulated with TNFalpha. Cells that were GFP+ after stimulation were collected for integration site sequencing. DNA was digested with MseI or a combination of NheI, SpeI and XbaI, ligated to adapters for nested PCR, amplified and sequenced by Sanger capillary electrophoresis.

Bcl-2 transduced $CD4^+$. Shan et al. [245] transduced $CD4^+$ T cells with Bcl-2, costimulated with bound anti-CD3 and soluble anti-CD28 antibodies, interleukin-2 and T cell growth factor and then infected with X4-pseudotyped GFP-expressing NL4-3-δ6-drEGFP construct [259] at an MOI of less than 0.1. DNA was extracted, digested with PstI and circularized [260]. HIV-human junctions were amplified by reverse PCR and sequenced using Sanger capillary electrophoresis.

Active $CD4^+$ & Resting $CD4^+$. Pace et al. [246] spinoculated $CD4^+$ T cells with HIV

NL4-3 at an MOI of 0.1. After 96 hours, the cells were stained for intracellular Gag CD25, CD69 and HLA-DR and sorted into four subpopulations based on activation state and Gag expression; activated Gag-, activated Gag+, resting Gag- and resting Gag+. The ability of the viruses to reactivate was not tested although previous studies have shown that the majority are likely inducible [261]. Genomic DNA was extracted and digested with restriction enzymes MseI and Tsp509 and ligated to adapters. Proviral LTR-host genome junctions were sequenced by 454 pyrosequencing after nested PCR.

All datasets were processed using the hiReadsProcessor R package [262]. Adaptor trimmed reads were aligned to UCSC freeze hg19 using BLAT [136]. Genomic alignments were scored and required to start within the first three bases of a read with 98% identity. Alignments for a given read with a BLAT score less than the maximum score for that read were discarded. Reads giving rise to multiple best scoring genomic alignments were excluded, while reads with a single best hit were dereplicated and converged if within 5bp of each other. The Bcl-2 transduced $CD4^+$ sample was sequenced from U3 in the $5'$ HIV LTR while the other samples were sequenced from U5 in the $3'$ LTR. To account for the 5 base duplication of host DNA caused by HIV integration, the chromosomal coordinates of the Bcl-2 transduced $CD4^+$ sample were adjusted by ± 4 bases.

To allow for alignment difficulties in the analysis of genomic repeats, reads with multiple best scoring alignments, along with the single best hit reads used above, were included in the repeat analyses. If any best scoring alignment for a read fell within a repeat, then that read was considered to map to that repeat.

3.3.4 Genomic features

A total of 140 whole genome features for $CD4^+$ T-cells were gathered from data sources indicated in Table 3.2. For features encoded as peaks or hotspots, the log of the distance of each integration site to the nearest border was used for modeling. Integration sites

Title	**Cell type**	**Virus**	**Time of harvest after infection**	**Sequencing**	**Generation of expressed vs. silent/ inducible**	**Citation**	**Silent/ inducible unique sites**	**Expressed unique sites**
Jurkat	Jurkat cells	HIV vector pEV731 (LTR-Tat-IRES-GFP)	2 weeks	Sanger	TNFα, GFP expression	Lewinski et al. [244]	463 inducible	643
Bcl-2 transduced $CD4^+$	Primary $CD4^+$ T cells (Bcl-2 transduced)	HIV NL4-3-δ6-drEGFP (inactivated *gag*, *vif*, *vpr*, *vpu*, *nef* and *env* replaced by GFP)	3 days + 3-4 weeks + 3 days	Sanger	anti-CD3, anti-CD28 antibodies, GFP expression	Shan et al. [245]	446 inducible	273
Active $CD4^+$	Primary active $CD4^+$ T cells	HIV NL4-3	3 days	454	high vs. low Gag	Pace et al. [246]	1604 silent	1274
Resting $CD4^+$	Primary resting $CD4^+$ T cells	HIV NL4-3	3 days	454	high vs. low Gag	Pace et al. [246]	1942 silent	784
Central Memory $CD4^+$	Primary central memory $CD4^+$ T cells	HIV NL4-3 ΔNef GFP	2 days/9 days	Ion-Torrent	anti-CD3, anti-CD28 antibodies, GFP expression	This paper	1729 inducible	3278

Table 3.1: HIV-1 integration datasets from *in vitro* models of latency where the proviruses were determined to be silent/inducible or expressed

from HIV 89.6 infection in primary $CD4^+$ T cells (unpublished data) were used to count nearby integrations and determine a ± 20bp position weight matrix for integration targets. Illumina RNA-Seq from active $CD4^+$ cells (unpublished data) was used to estimate raw cellular expression and fragments per kilobase of transcript per million mapped reads for genes as calculated by Cufflinks [108]. For sequence-based data like RNA-Seq and ChIP-Seq, the number of reads aligned within a $\pm$ 50, 500, 5,000 50,000 and 500,000 bp windows of each integration site were counted and log transformed. In addition, chromatin state classifications derived from a hidden Markov model based on histone marks and a few binding factors [263] were included as binary variables. All data from previous genomic freezes were converted to hg19 using liftover [264].

3.4 Results

The combination of integration site data newly reported here (set named "Central Memory $CD4^+$") with previously published data (sets named "Jurkat", "Bcl-2 transduced $CD4^+$", "Active $CD4^+$", and "Resting $CD4^+$") provides a collection of 12,436 integration sites (Table 3.1) where the expression status of the provirus—silent/inducible or expressed—is known. In three of the datasets, Jurkat, Central Memory $CD4^+$ and Bcl-2 transduced $CD4^+$, the proviruses were sorted based on inducibility. In the Resting $CD4^+$ and Active $CD4^+$ datasets, cells were sorted only based on proviral expression. Previous studies have shown that most silent proviruses in this model system are inducible [261].

3.4.1 Global model

If a genomic feature and latency are monotonically related then we should be able to detect this relationship using Spearman rank correlation. In addition if a feature has a consistent effect across models we should see a consistent pattern in the direction of correlation. A simple first look for correlation between genomic features (Table 3.2) and latency status yielded inconsistent results among the five samples with no variables

Group	Type	Source	Number	Types
T cell expression	RNA-Seq	Chapter 2	1	RNA
Jurkat expression	RNA-Seq	Encode [265]	1	wgEncodeHudsonalphaRnaSeq
Integration sites	Locations	Berry et al. [135]	1	sites
DNase sensitivity	DNA-Seq/peaks	Encode [265]	1	wgEncodeOpenChromDnase
Methylation	DNA-Seq	[266]	1	Methyl
CpG	Locations	UCSC [267]	1	cpgIslandExt
Sequence-based	Continuous	—	4	% GC, HIV PWM score, distance to centrosome, chromosomal position
Repeats	Locations	UCSC [267]	16	DNA, LINE, Low_complexity, LTR, Other, RC, RNA, rRNA, Satellite, scRNA, Simple_repeat, SINE, snRNA, srpRNA, tRNA, alphoid
Histone features	ChIP-Seq/Peaks	Wang et al. [268]	18	H2AK5ac, H2AK9ac, H2BK120ac, H2BK12ac, H2BK20ac, H2BK5ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K36ac, H3K4ac, H3K9ac, H4K12ac, H4K16ac, H4K5ac, H4K8ac, H4K91ac
Histone features	ChIP-Seq/Peaks	Barski et al. [269]	23	CTCF, H2AZ, H2BK5me1, H3K27me1, H3K27me2, H3K27me3, H3K36me1, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me1, H3K79me2, H3K79me3, H3K9me1, H3K9me2, H3K9me3, H3R2me1, H3R2me2, H4K20me1, H4K20me3, H4R3me2, PolII
Chromatin state	Binary	Ernst and Kellis [263]	51	$state_1$,$state_2$,...,$state_{51}$
HATs and HDACs	ChIP-Seq	Wang et al. [270]	11	Resting-HDAC1, Resting-HDAC2, Resting-HDAC3, Resting-HDAC6, Resting-p300, Resting-CBP, Resting-MOF, Resting-PCAF, Resting-Tip60, Active-HDAC6, Active-Tip60
Nucleosome	ChIP-Seq	Schones et al. [271]	2	Resting-Nucleosomes, Active Nucleosomes
UCSC genes	Locations	Hsu et al. [163]	4	in gene, in gene (same strand), gene count, distance to nearest gene, in exon, in intron

Table 3.2: Genomic data available for comparison to HIV integration sites

having a significant Spearman rank correlation across all, or even four out of five, of the samples (Figure 3.1). This suggests that there is not a consistent simple monotonic relationship between the genomic variable and latency, or that any such correlations are modest and not detectable across all studies given the available statistical power. We return to some of the stronger trends below.

To investigate whether a combination of variables may affect latency, we fit a lasso-regularized logistic regression, as implemented in the R package glmnet [272], to predict latency using the genomic variables. The relationship between silent/inducible status and each genomic variable was allowed to vary between models by including the interaction of genomic features with dummy variables indicating cellular model. The λ smoothing parameter of the lasso regression was optimized by finding the λ with lowest classification error in 480-fold cross validation and finding the simplest model with misclassification error within one standard error.

The proportion of silent/inducible sites varied between the samples. To avoid the model overfitting on this source of variation, an indicator variable for each sample was included in the base model. The base model with no genomic variables was selected as the best model by cross validation (Figure 3.2A). This suggest that there is not a consistent linear relationship between an additive combination of genomic variables and latency across all models.

When each dataset was fit individually with leave-one-out cross validation, improvements in cross-validated misclassification error were only observed in the Active $CD4^+$ (5.8% decrease in misclassification error, standard error: 2.1) and Jurkat (6.7% decrease in misclassification error, standard error: 3.5) samples (Figure 3.2B-F). There was no overlap in variables selected for the Active $CD4^+$ and Jurkat samples.

Finding little global association between latency and genomic features, we investigated whether predictors of latency reported previously by single studies were consistently

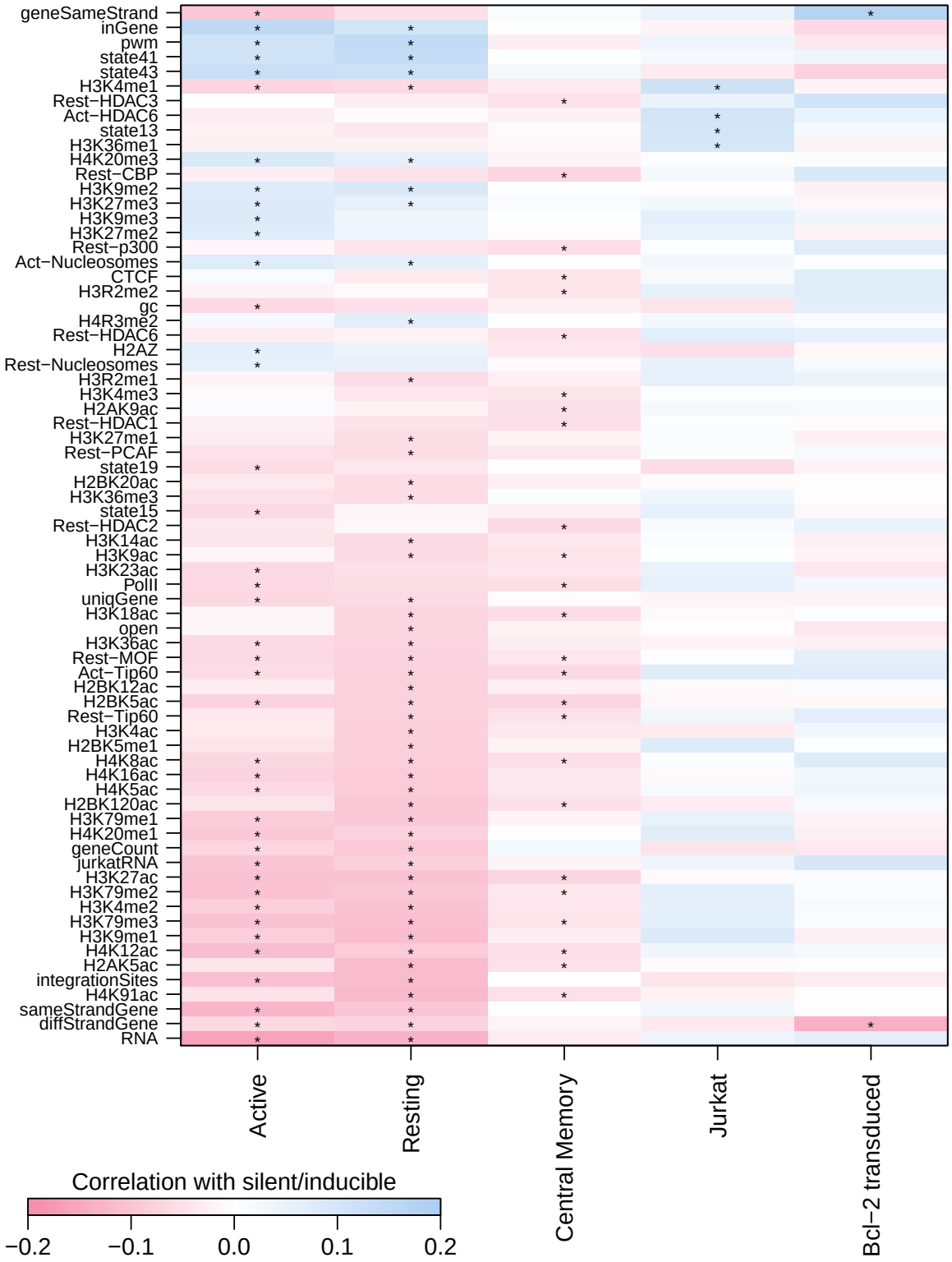


Figure 3.1: Spearman rank correlation between proviral expression status and genomic features. Only genomic features with at least one correlation with latency with a false discovery rate q-value < 0.01 (marked by asterisks) are shown.

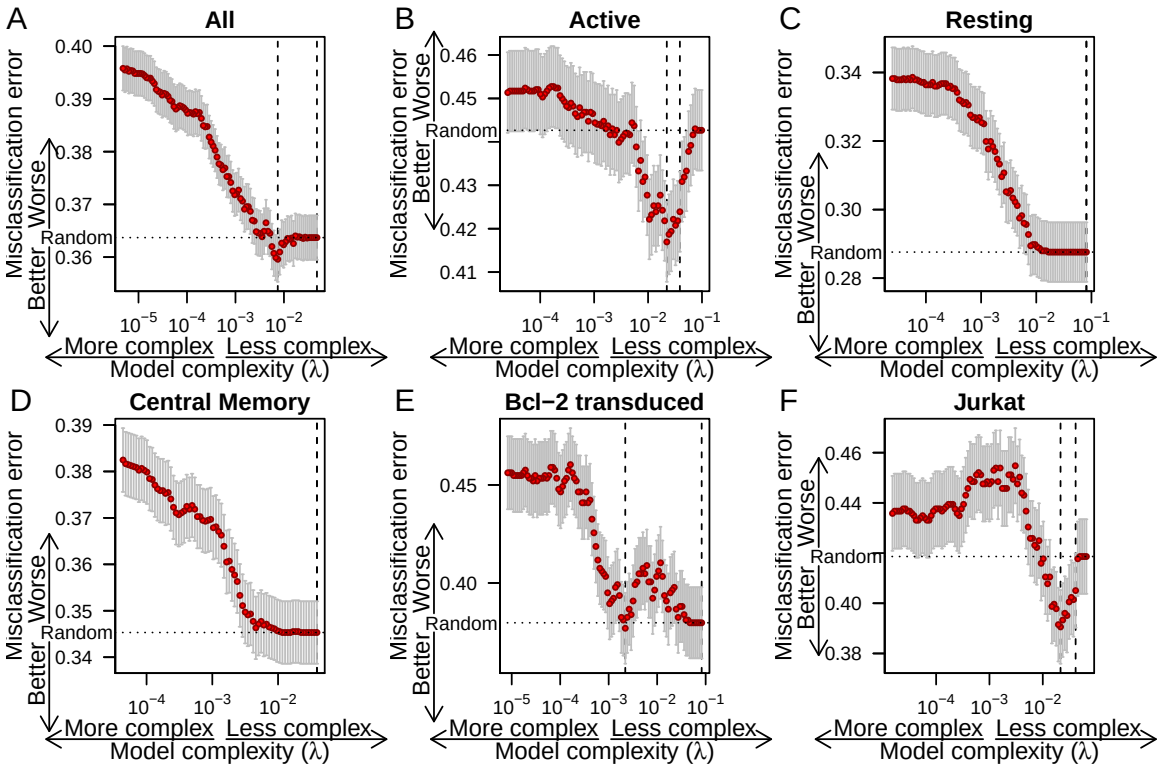


Figure 3.2: Misclassification error from cross validation for lasso regressions of silent/inducible status on genomic features as a function of λ, the regularization coefficient for the lasso regression, for all cell culture models combined and each individual cell culture model. The number of variables included and size of coefficients in the model increases to the left. Whiskers show the standard error of mean misclassification error. Dashed vertical lines indicate the minimum misclassification error and the simplest model within one standard error. Dotted horizontal line indicates the misclassification error expected from random guessing.

associated with latency across studies.

3.4.2 Cellular transcription

Model systems with defined integration sites show upstream transcription can interfere with viral transcription [273] and that cellular transcription in the same orientation may interfere with viral transcription [247] or increase viral transcription [248] and in opposite orientations may decrease transcription [248]. In integration site studies, integration outside genes appears to increase latency [244] but high transcription of nearby host cell genes may cause increased latency [244, 245]. In addition, Tat or other viral proteins may affect cellular transcription [70, 274].

To look at transcription and latency, we ran a logistic regression of silent/inducible status on a quartic function of RNA expression, as determined by RNA-Seq reads within 5,000 bases in Jurkat cells for the Jurkat sample or $CD4^+$ T cells for the remaining samples, interacted with indicator variables encoding cell culture model. There appears to be little agreement between samples (Figure 3.3). The Resting $CD4^+$ and Active $CD4^+$ datasets show an enrichment in silent proviruses in regions with low gene expression. The other three studies show the opposite or no relationship for low expression regions. The two samples showing increased silence in areas of low expression (Resting $CD4^+$ and Active $CD4^+$) are from a study that did not check whether inactive viruses could be activated. One possible explanation is that regions with low gene transcription may harbor proviruses that are not easily activated, though some other discrepancy between *in vitro* systems could also explain the difference. Both the Jurkat and Active $CD4^+$ samples appear to increase in latency with increasing expression while the remaining three studies did not show a strong trend.

3.4.3 Orientation bias

Shan et al. [245] reported that inducible proviruses were oriented in the same strand as the host cell genes into which they had integrated more often than chance. This

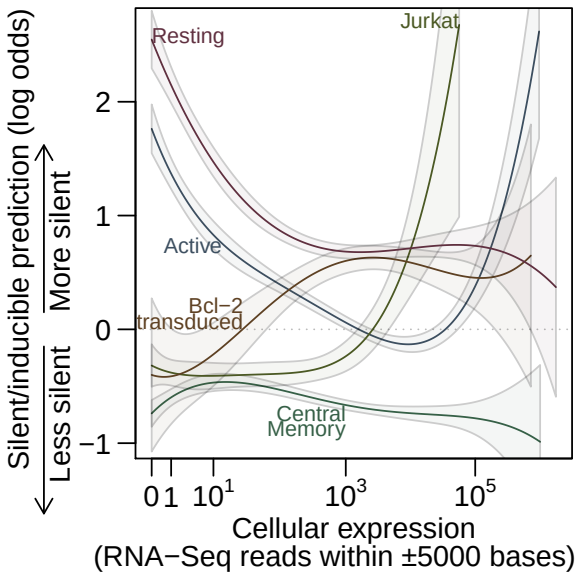


Figure 3.3: Predictions from a logistic regression of silent/inducible status on cellular RNA expression. High y-axis values are predicted to be silent/inducible. Dashed line shows where equal odds of silent/inducible and expressed are predicted. Solid lines show predictions from the regression for each sample and shaded regions indicate one standard error from the modeled predictions.

orientation bias was still reproduced after our reprocessing of the Bcl-2 transduced CD4$^+$ sample from Shan et al. [245]. However, the proportion of provirus oriented in the same strand as host genes did not differ significantly from 50% in the other samples (Figure 3.4.3). Perhaps orientation bias and transcriptional interference are especially sensitive to parameters of the model system.

3.4.4 Gene deserts

Lewinski et al. [244] reported increased latency in gene deserts. In the collected data, integration outside known genes was associated with latency (Fisher's exact test, $p < 10^{-6}$). This seemed to largely be driven by the Active CD4$^+$ and Resting CD4$^+$ samples with significant association found individually in only those two samples (both $p < 10^{-8}$) and no significant association observed in the other three samples (Figure 3.5A). Looking only at integration sites outside genes, silent sites in the Resting CD4$^+$ sample had a mean distance to the nearest gene 2.5 times greater than that of expressed sites (95% CI: 2.2–$6.2\times$, $p < 10^{-6}$, Welch two sample t-test on log transformed distance) (Figure 3.5B). The Active CD4$^+$ sample had a small difference that did not survive Bonferroni correction.

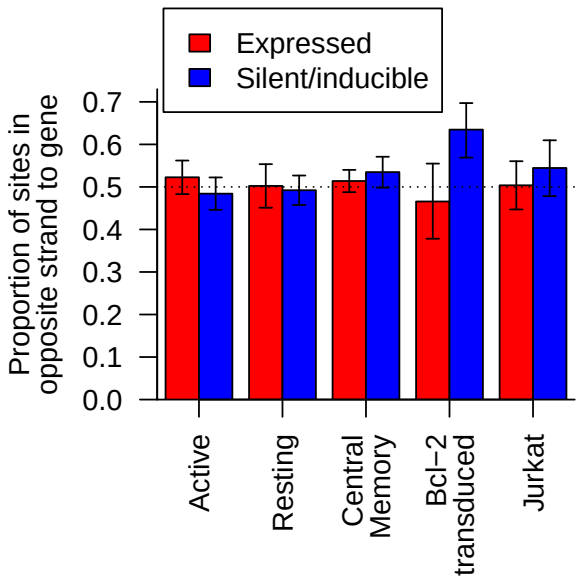


Figure 3.4: The proportion of provirus integrated in the opposite strand compared to cellular genes in silent/inducible (blue) and expressed (red) samples. Error bars show the 95% Clopper-Pearson binomial confidence interval.

Lewinski et al. [244] also reported decreased latency near CpG islands and reasoned this was tied to the increased latency in gene deserts. In the Resting $CD4^+$ sample, silent sites were on average further from CpG islands than expressed sites (Bonferroni corrected Welch's two sample T test, $p = 0.006$), but there was no significant relationship between silent/inducible status and log distance to CpG island after Bonferroni correction if the integration site's location inside or outside of a gene was accounted for first (analysis of deviance).

3.4.5 Alphoid repeats

Alphoid repeats are repetitive DNA sequences found largely in the heterochromatin of centromeres [275]. Integration near heterochromatic alphoid repeats has been reported to associate with latency [239, 244, 246]. Looking only at uniquely mapping sites, there was no statistically significant association between latency and location inside an alphoid repeat in pooled or individual samples (Fisher's exact test).

Since alphoid repeats are both problematic to assemble in genomes and difficult to map onto, we reasoned that some alphoid hits might be lost or miscounted in the filtering procedures of the standard workup. To counteract this, we treated each sequence read

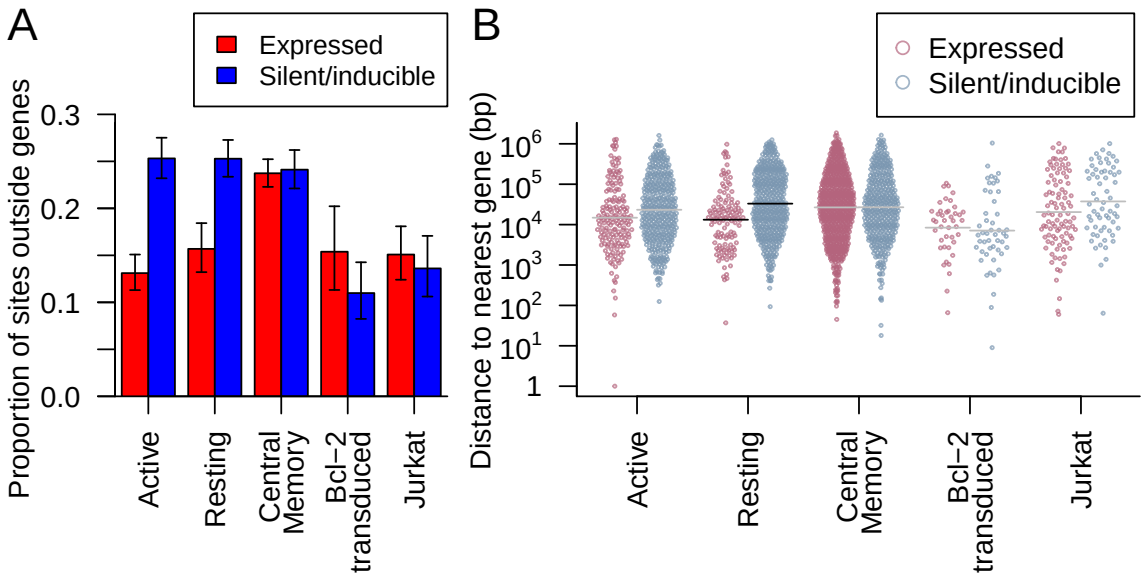


Figure 3.5: (A) The proportion of provirus integrated outside genes in silent/inducible (blue) and expressed (red) samples. Error bars show the 95% Clopper-Pearson binomial confidence interval. (B) The nearest distance to any gene for integration sites (points) outside genes in the five samples. Points are spread in proportion to kernel density estimates. Horizontal lines indicate sample means where there was a significant difference in means between silent/inducible and expressed provirus (black) or no significant difference (grey).

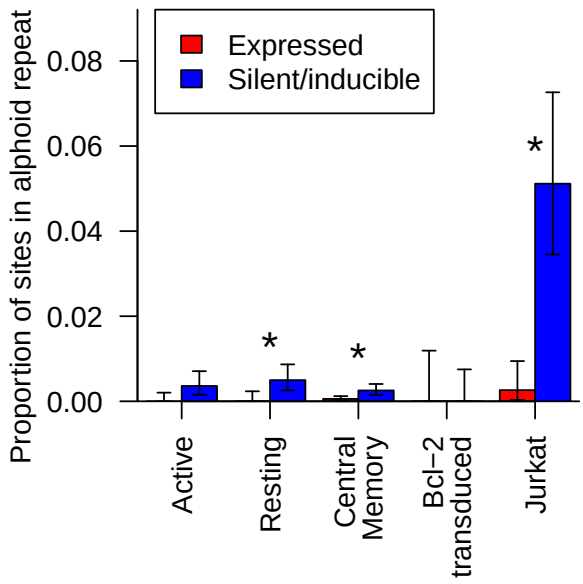


Figure 3.6: The proportion of integration sites with matches in alphoid repeats in silent/inducible (blue) and expressed (red) cells in five samples. Error bars show the 95% Clopper-Pearson binomial confidence interval. Asterisks indicate significant associations between integrations within an alphoid repeat and proviral expression status (Bonferroni corrected Fisher's exact test $p < 0.05$).

as an independent observation of a proviral integration and included sequence reads with more than one best scoring alignment. For multiply aligned reads, we considered the read to have been inside an alphoid repeat if any of its best scoring alignments fell within a repeat. We found 74 reads with potential alphoid mappings. Integration inside alphoid repeats was significantly associated with the expression status of a provirus in the Resting CD4$^+$, Jurkat and Central Memory CD4$^+$ datasets (Bonferroni corrected Fisher's exact test, all $p < 0.05$) and approached significance in the Active CD4$^+$ dataset ($p = 0.053$) (Figure 3.6). The Bcl-2 transduced CD4$^+$ data did not contain any integration sites in alphoid repeats, probably due to 1) the relatively low number of integration sites in the dataset and 2) to the requirement for cleavage at two Pst1 restriction sites, which are not found in the consensus sequence of alphoid repeats [164]. Of the 1340 repeat types in the RepeatMasker database [164], only alphoid repeats achieved a significant association with proviral expression in more than two datasets.

3.4.6 Acetylation

Histone marks or chromatin remodeling, especially involving the key "Nuc-1" histone near the transcription start site in the viral LTR, appear to affect viral expression [240, 276, 277]. Based on this effect, histone deacetylase inhibitors have been developed

as potential HIV treatments and show some promise in disrupting latency [252]. In these genome-wide datasets, we do not have information on the state of individual LTR nucleosomes. However, repressive chromatin does seem to spread to nearby locations if not blocked by insulators [236, 237] and the state of neighboring chromatin could affect proviral transcription independently of provirus-associated histones.

We found that the number of ChIP-seq reads near an integration site from several histone acetylation marks (Figure 3.1) were associated with efficient expression in the Active CD4$^+$, Resting CD4$^+$ and Central Memory CD4$^+$ samples. H4K12ac had the strongest association (Bonferroni corrected Fisher's method combination of Spearman's ρ, $p < 10^{-25}$) with silence/latency (Figure 3.7A).

Although the appearance of several significantly associated acetylation marks might suggest acetylation exerts a considerable effect on the expression of a provirus, there are strong correlations among these marks, so their effects may not be independent. To account for the correlations between these variables, we performed a principal component analysis (PCA) to convert the correlated acetylation marks into a series of uncorrelated principal components that capture much of the variance within a few components. Here, the first principal component explained 59% of the variance and the first ten components 84%. Several of these principal components again displayed significant associations with latency in the Active CD4$^+$, Resting CD4$^+$ and Central Memory CD4$^+$ samples but no significant correlations in the Bcl-2 transduced CD4$^+$ or Jurkat samples (Figure 3.7B). A logistic regression of expression status on the first ten principal components and sample did not reduce misclassification error from a base model including only sample in 480-fold cross validation (base model misclassification error: 36.4%, PCA model: 36.5%). This suggests that acetylation of neighboring chromatin does not exert strong effects on latency in all samples.

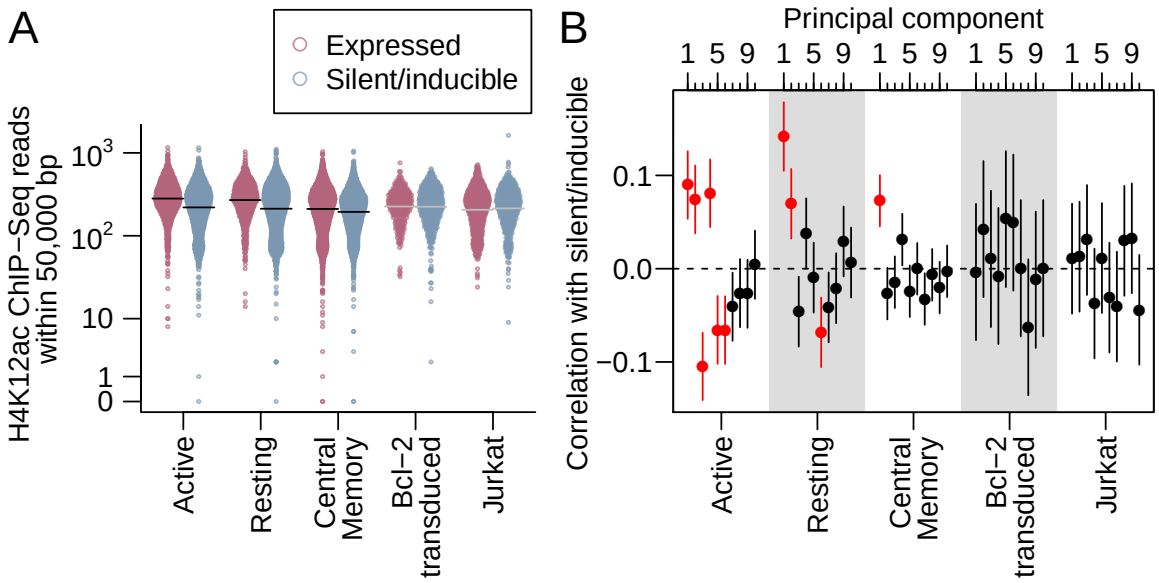


Figure 3.7: (A) The number of ChIP-seq reads for H4K12ac, the histone mark with the lowest Fisher's method p-value for correlation with latency, within 50,000 bases across the five samples. Integration sites (points) are spread in proportion to kernel density estimates. Horizontal lines indicate sample means where there was a significant difference (black) in means between silent/inducible and expressed provirus or no significant difference (grey). (B) The correlation (points) and its 95% confidence interval (vertical lines) between principal components of acetylation and silent/inducible status for each of the five samples. Red indicates correlations with a Bonferroni-corrected p-value < 0.05.

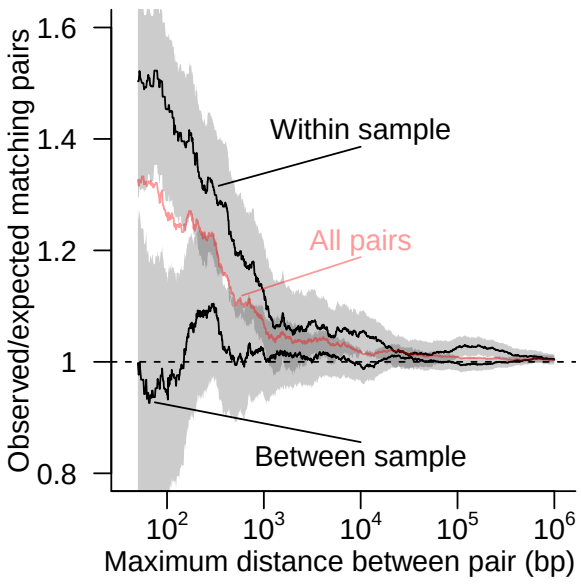


Figure 3.8: The ratio of the number of pairs of proviruses with matching expression status to the number of matches expected by random pairings given the frequency of silent/inducible proviruses. All possible pairs of proviruses integrated within a given distance of each other on the same chromosome (red line) were separated into two sets; one with both proviruses from within the same cell culture model and one with proviruses paired between two different cell culture models (black lines). The shaded region shows the 95% Clopper-Pearson binomial confidence interval for within and between sample pairings. The dashed horizontal line shows the ratio of 1 expected if there is no association between the expression status of neighboring proviruses.

3.4.7 Clustering

We reasoned that if there was a strong relationship between latency and chromosomal position, then integration sites that are near one another on the same chromosome should share the same expression status more often than expected by chance. To test this, we compared how often pairs of proviruses shared the same expression status in relation to the distance between the two sites (Figure 3.8). Pairs of sites with little distance between integration locations did share the same expression status more often than expected by chance (e.g. neighbors closer than 100bp, Fisher exact test $p = 0.0002$). Breaking out the data to separate between sample and within sample pairings showed that this matching was limited to neighbors within the same experimental model (Figure 3.8), emphasizing that chromosomal environment does appear to influence latency, but the factors involved differ among experimental models of latency.

3.5 Conclusions

Here we compared the latency status of HIV-1 proviruses in five model systems with the genomic features surrounding their integration sites. Surprisingly, no relationships between genomic features near the integration location and latency achieved significance in all models. Proviruses from the same cellular model integrated in nearby positions did share the same latency status much more often than predicted by chance, indicating the existence of local features influencing latency, but these were not consistent among models. This suggests that whatever features are affecting latency are highly local and model-specific, and that we may not have access to all relevant chromosomal features [e.g. 278–281].

In addition to differences in experimental conditions, methodological issues have the potential to obscure patterns. Examples include multiply infected cells, inactivated viruses and inaccurate assessment of HIV gene activity—each of these are discussed below.

A latent provirus integrated into the same cell as an expressed provirus will be erroneously sorted as expressed, potentially confounding analysis. A low multiplicity of infection (MOI) will help to avoid this problem, but there is still the potential for a significant proportion of the cells studied to contain multiple integrations. This problem arises because although cells with multiple integrations form a small proportion of total cells, most of the total are cells lacking an integrated provirus and thus are excluded by experimental design. For example, assuming integrations are Poisson distributed with an MOI of 0.1 (1 integration per 10 cells), 90.5% of cells will not contain a provirus, 9% of cells will contain one proviral integration and 0.5% of cells will contain multiple integrations. The cells without an integration are not amplified by HIV-targeted PCR leaving only 9.5% of the total cells. Of these cells actually under study, 4.9% will contain multiple integrations. Thus the signal from expressed proviruses may be muted by the presence of latent proviruses in the expressed population.

The replication cycle of HIV is error prone, and a significant proportion of virions contain mutated genomes [282]. In studies that do not check for inducibility, mutant proviruses integrated in regions of the genome otherwise favorable to proviral expression can be sorted into the latent pool due to mutational inactivation. This problem of inactivated provirus is worse when latent provirus are rare and exacerbated further when looking at latency in the cells of HIV patients due to selective enrichment of inactivated proviruses incapable of spreading infection [227]. Here, the effects of mutation are minimized in the datasets that required inducible viral expression (Jurkat, Bcl-2 transduced $CD4^+$, Central Memory $CD4^+$) but may be a confounder in the two datasets that were sorted based on lack of viral expression only (Active $CD4^+$, Resting $CD4^+$).

Inaccurate staining or leaky markers may also result in misclassification of proviruses. False positives and false negatives will result in incorrectly sorted latent and expressed integrations. For example, if 5% of cells not containing Gag are labeled as Gag+ and there are an equal amount of latent and expressed integration sites, then 4.8% of integrations labeled expressed will actually be latent. If a category is rare, false staining has even greater potential to cause error. For example, if only 5% of sites are latent and a Gag stain has a false negative rate of 5%, then we would expect 48.7% of sites classified as latent to actually be mislabeled expressed integrations.

Attempts to induce latent proviruses in patients have so far focused on using histone deacetylase inhibitors, raising interest in associations with histone acetylation in these data. An important caveat in results from these genome-wide data is that histone modification near the integrated provirus may not be representative of modification within the provirus at the key "Nuc-1" nucleosome of the transcription start site [277], though local correlations in chromatin states are well established from studies of position effect variegation [236, 237]. We found that some histone acetylation marks were significantly associated with viral expression in some but not all samples (Figures 3.1, 3.7). This lack of association may be due to a lack of power in these studies, but

the confidence intervals suggest that any correlations between acetylations and latency are unlikely to be strong. These weak correlations raise the possibility that there are populations of latent proviruses that are not associated with acetylation and may not be inducible by histone deacetylase inhibitors.

This study highlights that the choice of model system can have a large effect on measurements of latency. Further studies are needed to determine which *in vitro* models best reflect latency *in vivo*. Different cell models may report genuinely different mechanisms of latency. While we did see some relationship between histone acetylation and latency, paralleling a recent clinical trial of SAHA [252], associations with histone acetylation did not explain a large fraction of the difference between latent and expresssed proviruses in any of the five models. One possible explanation is that there may be multiple mechanisms that maintain proviruses in a latent state. To be successful, shock-and-kill treatments must induce and destroy all latent proviruses to eliminate HIV from an infected individual, raising the question of whether multiple simultaneous inducing treatments will be necessary.

3.6 Availability of supporting data

Sequence reads from the Central Memory $CD4^+$ sample reported here, the Resting $CD4^+$ and Active $CD4^+$ data reported by Pace et al. [246], the Bcl-2 transduced $CD4^+$ data reported by Shan et al. [245] and reprocessed data originally reported by Lewinski et al. [244] are available at the Sequence Read Archive under accession number SRP028573.

3.7 Author's contributions

SS-M led the computational analysis, with assistance from CCB and NM. MKL, DL and JG analyzed integration sites using IonTorrent sequencing. MF, AB and VP prepared DNA from latent and activated T cells using the Central Memory $CD4^+$ model. LS, RFS, MJP, LMA and UO'D contributed data and suggestions. SS-M, KEO and FDB planned the overall study, and SS-M and FDB wrote the paper. All authors read and

approved the final manuscript.

3.8 Acknowledgements

We would like to thank Werner Witke for assistance with IonTorrent sequencing. This work was supported in part by NIH grants R01 AI 052845-11 to FDB, R21AI 096993 and K02AI078766 to UO'D, 5T32HG000046 to SS-M, AI087508 to VP and R01AI038201 to JG, the Penn Genome Frontiers Institute, the University of Pennsylvania Center for AIDS Research (CFAR) P30 AI 045008 and the University of California, San Diego, CFAR P30 AI036214.

CHAPTER 4 : A reverse transcription loop-mediated isothermal amplification assay optimized to detect multiple HIV subtypes

4.1 Abstract

Diagnostic methods for detecting and quantifying HIV RNA have been improving, but efficient methods for point-of-care analysis are still needed, particularly for applications in resource-limited settings. Detection based on reverse-transcription loop-mediated isothermal amplification (RT-LAMP) is particularly useful for this, because when combined with fluorescence-based DNA detection, RT-LAMP can be implemented with minimal equipment and expense. Assays have been developed to detect HIV RNA with RT-LAMP, but existing methods detect only a limited subset of HIV subtypes. Here we report a bioinformatic study to develop optimized primers, followed by empirical testing of 44 new primer designs. One primer set (ACeIN-26), targeting the HIV integrase coding region, consistently detected subtypes A, B, C, D, and G. The assay was sensitive to at least 5000 copies per reaction for subtypes A, B, C, D, and G, with Z-factors of above 0.69 (detection of the minor subtype F was found to be unreliable). There are already rapid and efficient assays available for detecting HIV infection in a binary yes/no format, but the rapid RT-LAMP assay described here has additional uses, including 1) tracking response to medication by comparing longitudinal values for a subject, 2) detecting of infection in neonates unimpeded by the presence of maternal antibody, and 3) detecting infection prior to seroconversion.

4.2 Introduction

Despite the introduction of efficient antiretroviral therapy, HIV infection and AIDS continue to cause a worldwide health crisis [283]. Methods for detecting HIV infection have improved greatly with time [284]–today rapid assays are available that can detect HIV infection in a yes-no format using a home test kit that detects antibodies in saliva. Viral load assays that quantify viral RNA with quick turn-around time are widely

available in the developed world. However, quantitative viral load assays are not commonly available with actionable time scales in much of the developing world. This motivates the development of new rapid and quantitative assays that can be used at the point of care with minimal infrastructure [285, 286].

One simple and quantitative detection method involves reverse transcription-based loop mediated isothermal amplification (RT-LAMP) [287]. In this method, a DNA copy of the viral RNA is generated by reverse transcriptase, and then isothermal amplification is carried out to increase the amount of total DNA. Primer binding sites are chosen so that a series of strand displacement steps allow continuous synthesis of DNA without requiring thermocycling. Reaction products can be detected by adding an intercalating dye to reaction mixtures that fluoresces only when bound to DNA, allowing quantification of product formation by measurement of fluorescence intensity. Such assays can potentially be packaged in simple self-contained devises and read out with no technology beyond a cell phone.

RT-LAMP assays for HIV-1 have been developed previously and reported to show high sensitivity and specificity for subtype B, the most common HIV strain in the developed world [286, 288, 289]. Another recent study reported RT-LAMP primer set optimized for the detection of HIV variants circulating in China [290], and another on confirmatory RT-LAMP for group M viruses [291]. Assays have also been developed for HIV-2 [292]. A complication arises in using available RT-LAMP assays due to the variation of HIV genomic sequences among the HIV subtypes [293, 294], so that an RT-LAMP assay optimized for one viral subtype may not detect viral RNA of another subtype [295]. Tests presented below show that many RT-LAMP assays are efficient for detecting subtype B, for which they were designed, but often performed poorly on other subtypes. Subtype C infects the greatest number of people worldwide, including in Sub-Saharan Africa, where such RT-LAMP assays would be most valuable, motivating optimization for subtype C. Several additional non-B subtypes are also responsible for significant

burdens of disease world-wide [296].

Here we present the development of an RT-LAMP assay capable of detecting HIV-1 subtypes A, B, C, D, and G. We first carried out a bioinformatic analysis to identify regions conserved in all the HIV subtypes. We then tested 44 different combinations of RT-LAMP primers targeting this region in over 700 individual assays, allowing identification of a primer set (ACeIN-26) that was suitable for detecting these subtypes. We propose that the optimized RT-LAMP assay may be useful for quantifying HIV RNA copy numbers in point-of-care applications in the developing world, where multiple different subtypes may be encountered.

4.3 Methods

4.3.1 Viral strains used in this study

Viral strains tested included HIV-1 92/UG/029 (Uganda) (subtype A, NIH AIDS Reagent program reagent number 1650), HIV-1 THRO (subtype B, plasmid derived, University of Pennsylvania CFAR) [297], CH269 (subtype C, plasmid derived, University of Pennsylvania CFAR) [297], UG0242 (subtype D, University of Pennsylvania CFAR), 93BRO20 (subtype F, University of Pennsylvania CFAR), HIV-1 G3 (subtype G, NIH AIDS Reagent program reagent number 3187) [298].

Viral stocks were prepared by transfection and infection. Culture supernatants were cleared of cellular debris by centrifugation at 1500g for 10 min. The supernatant containing virus was then treated with 100 U DNase (Roche) per 450 uL virus for 15 min at 30°C. RNA was isolated using the QiaAmp Viral RNA mini kit (Qiagen GmbH, Hilden, Germany). RNA was eluted in 80 uL of the provided elution buffer and stored at -80°C.

Concentration of viral RNA copies was calculated from p24 capsid antigen capture assay results provided by the University of Pennsylvania CFAR or the NIH AIDS-

reagent program. In calculating viral RNA copy numbers, we assumed that all p24 was incorporated in virions, all RNA was recovered completely from stocks, 2 genomes were present per virion, 2000 p24 molecules per viral particle, and the molecular weight of HIV-1 p24 was 25.6 kDa.

4.3.2 Assays

RT-LAMP reaction mixtures (15 μL) contained 0.2 μM each of primers F3 and B3 (if a primer set used multiple B3 primers, mixture contained 0.2 μM of each); 1.6 μM each of FIP and BIP primers (if a primer set had multiple FIP primes, reaction mixture contained 0.8 μM of each FIP primer); and 0.8 μM each of LoopF and LoopB primers; 7.5 μL OptiGene Isothermal Mastermix ISO-100nd (Optigene, UK), ROX reference dye (0.15 μL from a 50X stock), EvaGreen dye (0.4 μL from a 20X stock; Biotium, Hayward, CA); HIV RNA in 4.7 μL; AMV reverse transcriptase (10U/μL) 0.1 μL and water to 15 μL In most cases where two primer sets were combined, the total primer concentration within the reaction was doubled such that the above individual primer molarities were maintained. For the mixture ACeIN-26+F-IN (S2 Table, line 46), the total primer concentration was not doubled–the F-IN primer set comprised 25% of the total primer concentration, and the ACeIN-26 primer set comprised 75% of the total primer concentration with the ratios of primers listed above preserved. This mixture was combined 1:1 with the ACe-PR primer set (S2 Table, line 47) such that total primer concentration in the final mixture was doubled.

Amplification was measured using the 7500-Fast Real Time PCR system from Applied Biosystems with the following settings: 1 minute at 62°C; 60 cycles of 30 seconds at 62°C and 30 seconds at 63°C. Data was collected every minute. Product structure was assessed using dissociation curves which showed denaturation at 83°C. Products from selected amplification reactions were analyzed by agarose gel electrophoresis and showed a ladder of low molecular weight products (data not shown).

Product synthesis was quantified as the cycle of threshold for 10% amplification. Z-factors [299] were calculated from tests of 24 replicates using the ACeIN26 primer set in assays with viral RNA of each subtype. No detection after 60 min was given a value of 61 min in the Z-factor calculation.

4.4 Results

4.4.1 Testing published RT-LAMP primer sets against multiple HIV subtypes

We first assessed the performance of existing RT-LAMP assays on RNA samples from multiple HIV subtypes. We obtained viral stocks from HIV subtypes A, B, C, D, F, and G, estimated the numbers of virions per ml, and extracted RNA. RNAs were mixed with RT-LAMP reagents which included the six RT-LAMP primers, designated F3, B3, FIP, BIP, LF and LB [287]. Reactions also contained reverse transcriptase, DNA polymerase, nucleotides and the intercalating fluorescent EvaGreen dye, which yields a fluorescent signal upon DNA binding. DNA synthesis was quantified as the increase in fluorescence intensity over time, which yielded a typical curve describing exponential growth with saturation (examples are shown below). Results are expressed as threshold times (T_t) for achieving 10% of maximum fluorescence intensity at the HIV RNA template copy number tested.

In initial tests, published primer sets targeting the HIV-1 subtype B coding regions for capsid (CA), protease (PR), and reverse transcriptase (RT) (named B-CA, B-PR and B-RT) were assayed in reactions with RNAs from four of the subtypes. Results with each primer set tested are shown in Figure 4.1 in heat map format, where each tile summarizes the results of tests of 5000 RNA copies. Primers and their groupings into sets are summarized in S1 and S2 Tables, average assay results are in S3 Table, and raw assay data is in S4 Table. Assays (Figure 4.1, top) with the B-CA, B-PR and B-RT primer sets detected subtypes B and D at 5000 RNA copies with threshold times less than 20 min. However, assays with B-CA and B-RT detected subtypes C and F with

threshold times > 50 min, indicating inefficient amplification and the potential for poor separation between signal and noise. B-PR did not detect subtype C at all. In an effort to improve the breadth of detection, we first tried mixing the B-PR primers, which detected clade F (albeit with limited efficiency) with the B-CA and B-RT primers (Figure 4.1 and S3 and S4 Tables). In neither case did this provide coverage of all four clades tested. We thus did not test these primer sets on RNAs from the remaining subtypes and instead sought to develop primer sets targeting different regions of the HIV genome.

4.4.2 Primer design strategy

To design primers that detected multiple HIV subtypes efficiently, we analyzed alignments of HIV genomes (downloaded from the Los Alamos National Laboratory site [293]) for regions with similarity across most viruses, revealing that a segment of the pol gene encoding IN was particularly conserved (Figure 4.2A). A total of six primers are required for each RT-LAMP assay [287]. We used the EIKEN primer design tool to identify an initial primer set targeting this region. In further analysis, positions in the alignments were identified within primer landing sites that commonly contained multiple different bases. Primer positions were manually adjusted to avoid these bases when possible, and when necessary mixtures were formulated containing each of these commonly occurring bases (S1 and S2 Tables). An extensive series of variants targeting the IN coding region was tested empirically in assays containing RNAs from multiple subtypes (5000 RNA copies per reaction, over 700 total assays; S3 and S4 Tables). Based on initial results, primers were further modified by adjusting the primer position or addition of locked nucleic acids as described below.

4.5 Testing different primer designs

Our first design, ACeIN-1 (“ACe” for “All Clade” and “IN” for “integrase”), targeted the HIV IN coding region and contained multiple bases at selected sites to broaden

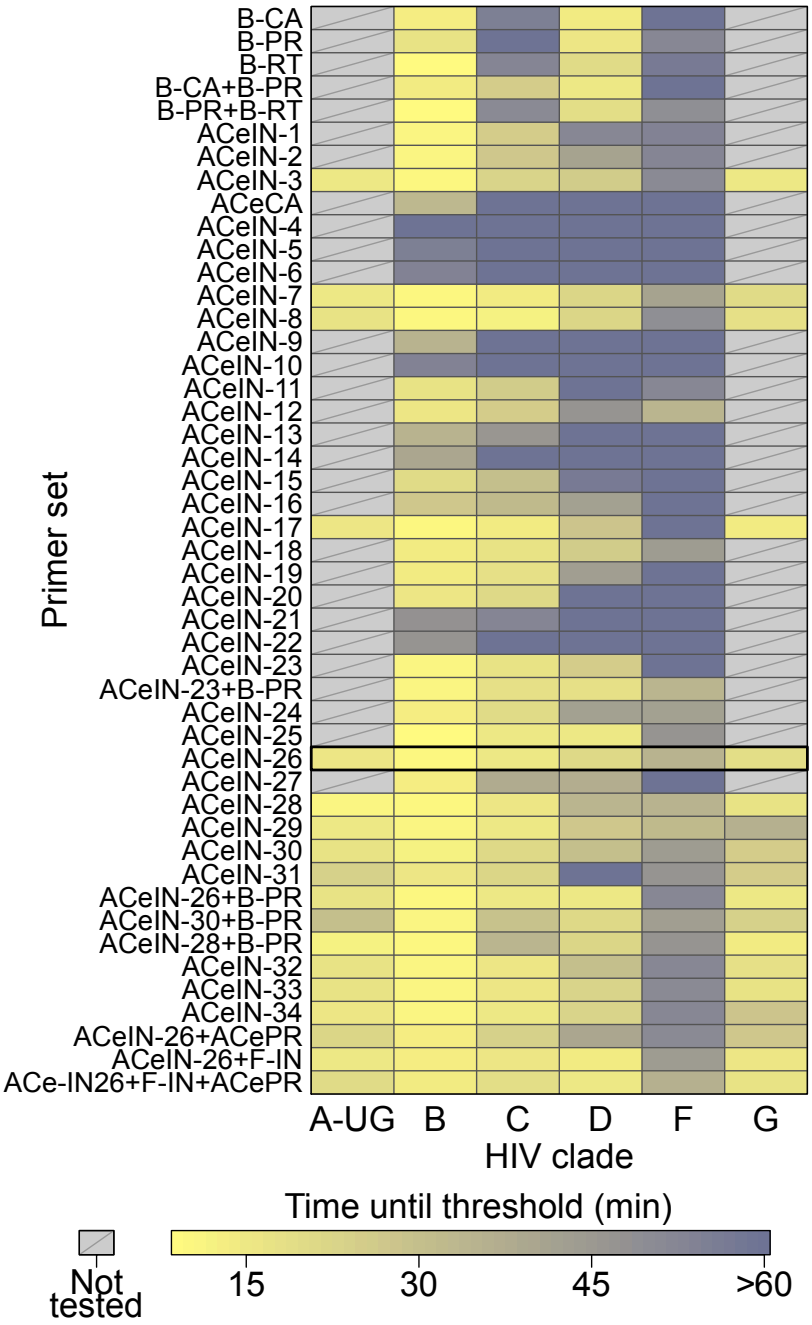


Figure 4.1: Summary of amplification results for all the RT-LAMP primer sets tested in this study. The data is shown as a heat map, with more intense yellow coloring indicating shorter amplification times (key at bottom). Primer sets tested are named along the left of the figure. Primer sequences, and their organization into LAMP primer sets, are cataloged in S1 and S2 Tables. The raw data and averaged data are collected in S3 and S4 Tables. ACeIN-26 primer set (highlighted) had one of the best performances across the subtypes and a relatively simple primer design.

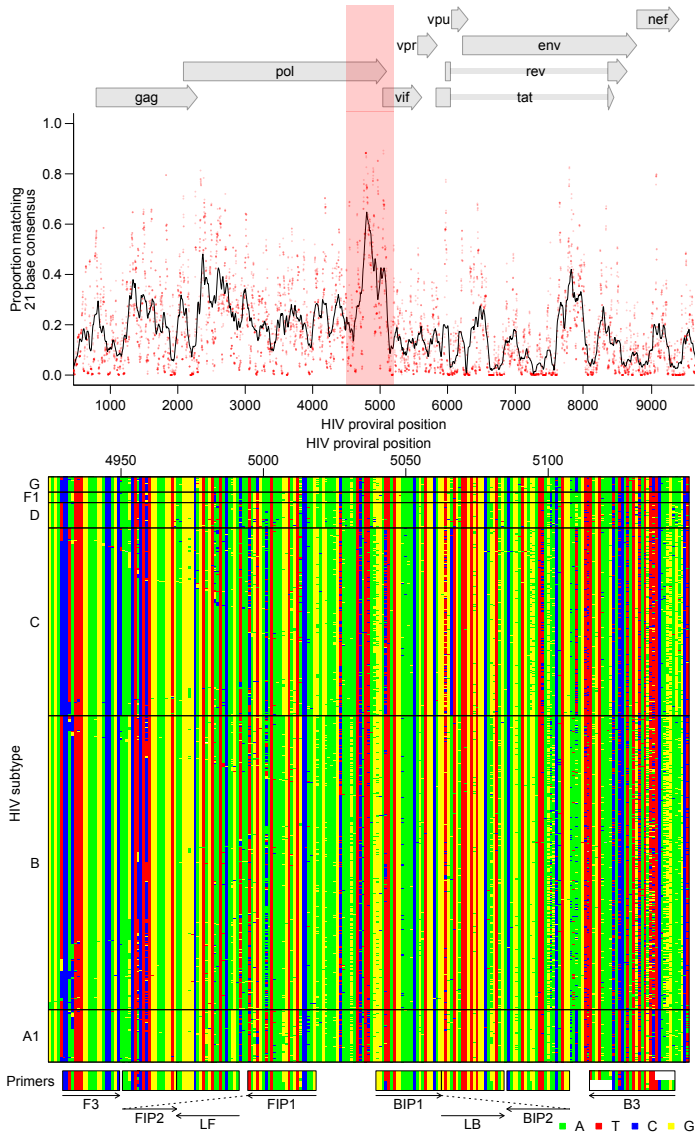


Figure 4.2: Bioinformatic analysis to design subtype-agnostic RT-LAMP primers. A) Conservation of sequence in HIV. HIV genomes (n = 1340) from the Los Alamos National Laboratory collection were aligned and conservation calculated. The x-axis shows the coordinate on the HIV genome, the y-axis shows the proportion of sequences matching the consensus for each 21 base segment of the genome (red points). The black line shows a 101 base sliding average over these proportions. The vertical red shading shows the region targeted for LAMP primer design that was used as input into the EIKEN primer design tool. Numbering is relative to the $HIV_{89.6}$ sequence. B) Aligned genomes, showing the locations of the ACeIN-26 primers. Sequences in the red shaded region in A are shown, with DNA bases color-coded as shown at the lower right. Each row indicates an HIV sequence and each column a base in that sequence. Horizontal lines separate the HIV subtypes (labeled at right). Arrows indicate the strand targeted by each primer. Primers targeting the negative strand of the virus are shown as reverse compliments for ease of viewing.

detection (Figure 4.1). ACeIN-2 and-3 have primers (B3) with slightly different landing sites. Tests showed that the mixture of primers allowed amplification with a shorter threshold time than did either alone (Figure 4.1).

We also tried to design a new primer set to the CA coding region (Figure 4.1, ACeCA) but found that the set only amplified clade B, and not efficiently. Thus this design was abandoned.

ACeIN-3 through-6 were altered by inserting a polyT sequence between the two different sections of FIP and BIP in various combinations, a modification introduced with the goal of improving primer folding, but these designs performed quite poorly (Figure 4.1).

Because the FIP primer appeared to bind the region with most variability among clades, we tried variations that bound to several nearby regions. These were tried with and without the polyT containing BIP and FIP primers in various combinations (Figure 4.1, ACeIN-7 through-22). We also tried mixing all of the variations of FIP together (ACeIN-23; S2 Table). The ACeIN-23 primer set was tried as a mixture with the B-PR set to try to capture clade F, yielding a relatively effective primer set (Figure 4.1, ACeIN-23+B-PR).

In an effort to increase affinity, an additional G/C pair was added to F3 and tested with various other IN primers (Figure 4.1, ACeIN-24 through-31). Testing showed improvement, with ACeIN-26 showing particularly robust amplification.

In a second effort to increase primer affinities, we substituted locked nucleic acids (LNAs) for selected bases that were particularly highly conserved among subtypes (Figure 4.1, ACeIN-30, -31, -32, -33, and-34). Some improvement was shown over the non-LNA containing bases. However, the ACeIN-26 primer set was as effective as or better than any LNA containing primer sets.

In further tests, the ACeIN-26, -28 and-30 primers were tested combined with the

ACePR primer set (a slightly modified version of the B-PR primer set, S2 Table, row 2, designed to accommodate a wider selection of HIV-1 subtypes) but no improvement was seen and efficiency may even have fallen for some subtypes. We also designed a primer set that matched exactly to the targeted sequences found in the problematic subtype F, and mixed this set with the ACeIN-26 primers. However, no improvement was seen (Figure 4.1, mixtures with F-IN set). Mixing the ACeIN-26 primers with both the ACePR and F-specific primers did yield effective primer sets (Figure 4.1, ACeIN26+F-IN and ACeIN26+F-IN+ACePR). However, amplification efficiency was not greatly improved over the ACeIN-26 primer set, so we proceeded with the simpler ACeIN-26 primer set (Figure 4.2B) in further studies.

4.5.1 Performance of the optimized RT-LAMP assay

The ACeIN-26 RT-LAMP primer set was next tested to determine the minimum concentration of RNA detectable under the reaction conditions studied (Figure 4.3). RNA template amounts were titrated and time to detection quantified. Tests showed detection after less than 20 min of incubation for 50 copies of subtypes A or B, detection after less than 30 min for 5000 copies for C, D, and G, and detection after less than 20 min for 50,000 copies for F.

For clinical implementation the reliability of an assay is critical. This is commonly summarized as a Z-factor [299], which takes into account both the separation in means between positive and negative samples and the variance in measurement of each. An assay with a Z-factor above 0.5 is judged to be an excellent assay. Z-factors for detection of each of the subtypes at 5000 RNA copies per reaction were > 0.50 for subtypes A, B, C, D, and G, respectively (Figure 4.4, n = 24 replicates per test). Detection of subtype F at 5000 copies per reaction was sporadic, showing a much lower Z-factor. Therefore our ACeIN-26 RT-LAMP primer set appears well suited to detect 5000 copies of subtypes A, B, C, D and G.

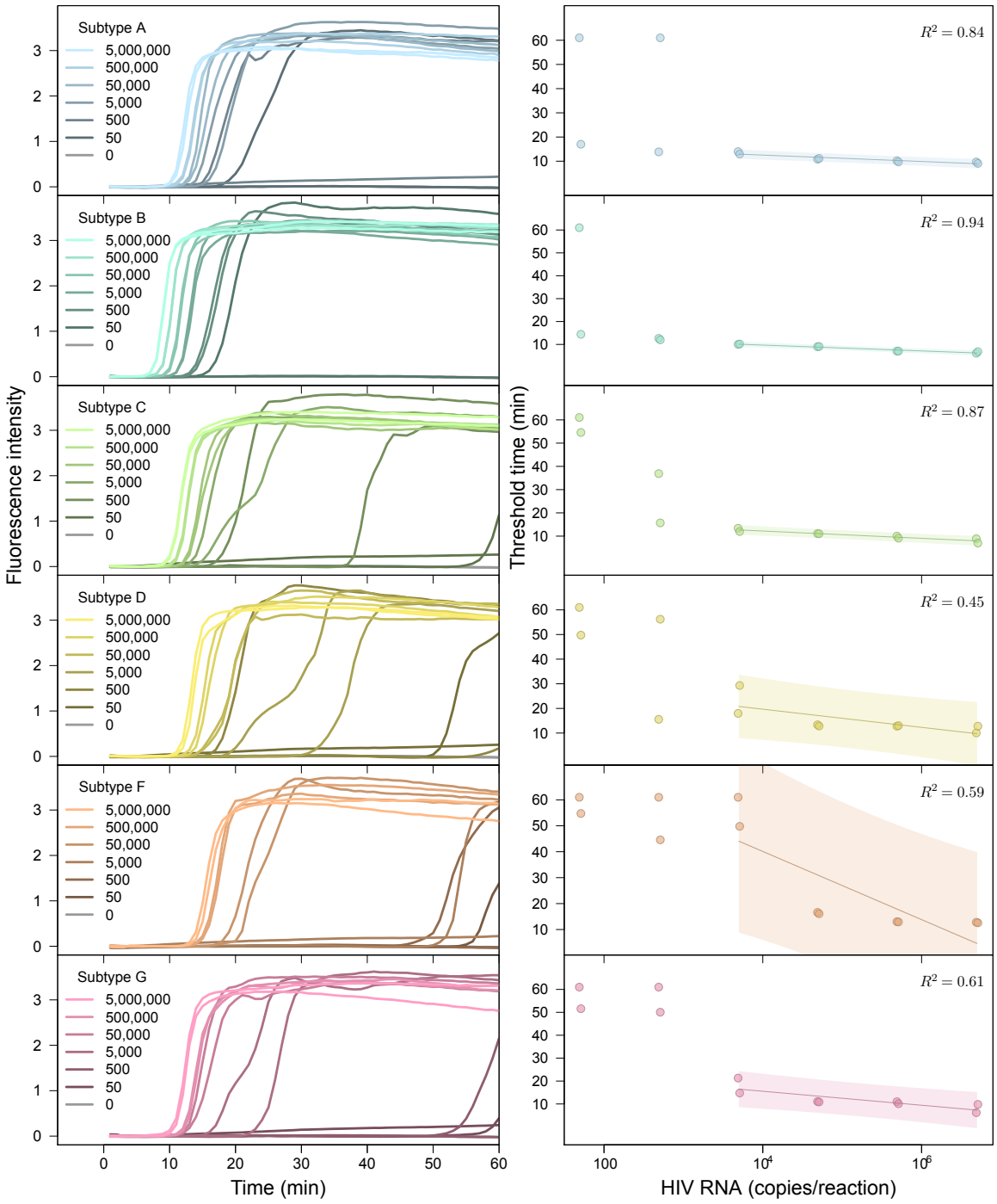


Figure 4.3: Performance of the AceIN-26 primer set with different starting RNA concentrations. Tests of each subtype are shown as rows. In each lettered panel, the left shows the raw accumulation of fluorescence signal (y-axis) as a function of time (x-axis); the right panel shows the threshold time (y-axis) as a function of log RNA copy number (x-axis) added to the reaction. In the right hand panels, values were dithered where two points overlapped to allow visualization of both.

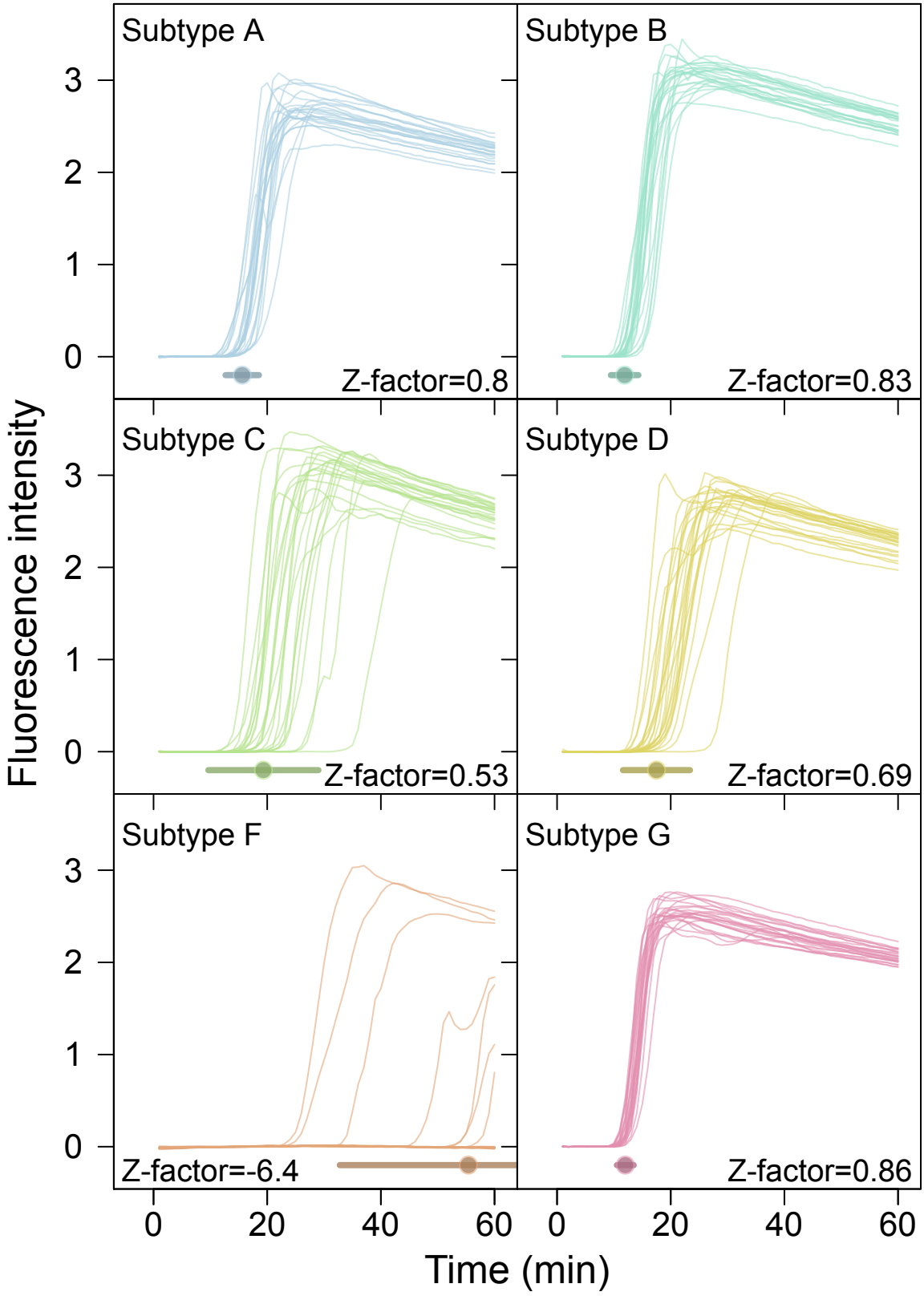


Figure 4.4: Examples of time course assays, displaying replicate tests of RT-LAMP primer set ACeIN-26 tested over six HIV subtypes, used in Z-factor calculations. A total of 5000 RNA copies were tested in each 15μL reaction. Time is shown on the x-axis, Fluorescence intensity on the y-axis. Replicates are distinguished using an arbitrary color code. Z-factor values and standard deviations are shown on each panel.

4.6 Discussion

Here we present an RT-LAMP assay optimized to identify multiple HIV subtypes. Infections with subtype B predominate in most parts of the developed world, but elsewhere other subtypes are more common [296]. Thus nucleic acid-based assays for use in the developing world need to query HIV subtypes more broadly. Previously reported RT-LAMP assays, while effective at detecting subtype B, commonly showed poor ability to detect at least some of the HIV subtypes, including C, which is common in the developing world (Figure 4.1). Here we first carried out an initial bioinformatic survey to identify regions conserved across all HIV subtypes that could serve as binding sites for RT-LAMP primers. We then tested primer sets targeting these regions empirically for efficiency. Testing 44 different primer sets revealed that assays containing ACeIN-26 were effective in detecting 5000 copies of RNA from subtypes A, B, C, D, and G within 30 minutes of incubation. For these five subtypes, the times of incubation to reach the threshold times were not too different, which simplifies interpretation when the subtype in the sample is unknown. Regardless of the efficiency, these assays can be applied to longitudinal studies of changes in viral load within an individual. We propose that RT-LAMP assays based on the ACeIN-26 primer set can be useful world-wide for assaying HIV-1 viral loads in infected patients.

There are several limitations to our study. Subtypes A, B, C, D, and G were detected efficiently and showed Z-factors above 0.5, but subtype F was detected reliably only with higher template amounts, probably due to more extensive mismatches with the ACeIN-26 primer set. Subtype F is estimated, however, to comprise only 0.59% of all infections globally [296], though it is common in some regions. For many of the common circulating recombinant forms, such as AE and BC, the target site for ACeIN-26 is from a subtype known to be efficiently detected, though in some cases the efficiency of detection is not easy to predict and will need to be tested. We did not test subtypes beyond A, B, C, D, F and G, and we did not attempt to assess multiple different variants

within each subtype. Thus, while we do know that our RT-LAMP assays are more widely applicable than many of those reported previously, we do not know whether they are able to detect all strains efficiently. In addition, although we carried out more than 700 assays in this study, there remain multiple parameters that could be optimized further, such as primer concentrations, salt type and concentration, temperature, and divalent metal concentrations, so there are likely further opportunities for improvement. Also, possible effects of RNA quality on assay performance were not tested rigorously.

A particularly important parameter for further optimization is primer sequence. Several groups have recently published primer sets optimized for broad detection of different HIV lineages [290, 291], offering opportunities for creating sophisticated primer blends with increased breadth of detection. However, in developing such mixtures, it will be important to monitor for possible complicating interactions of primers with each other. As an example of ongoing development of mixtures, we found that addition of another primer to the ACeIN-26 set that was matched to a common subtype C lineage allowed improved detection of subtype C variants (S1 Report). In order to improve detection of subtype F, which was suboptimal with ACeIN-26, additional primer sets could be mixed to specifically target subtype F, though the ones we tried so far did not work well. It will be useful to explore the performance of broader primer mixtures in future work.

Today rapid assays are available that can report infection efficiently, for example by detecting anti-HIV antibodies in oral samples–however, the nucleic acid-based method presented here has additional potential uses. We envision combining the RT-LAMP assay with simple point-of-care devices for purifying blood plasma [285] and quantitative analysis of accumulation of fluorescent signals [300]. In one implementation of the technology, cell phones could be used to capture and analyze results, thereby minimizing equipment costs. Point-of-care devices are available facilitating the concentration of viral RNA from blood plasma or saliva [300] to allow the detection of the 1000 RNA copy threshold that the WHO defines as virological treatment failure (World Health

Organization, Consolidated ARV guidelines, June 2013). Together, these methods will allow assessment of parameters beyond just the presence/absence of infection. Quantitative RT-LAMP assays should allow tracking of responses to medication, detection in neonates (where immunological tests are confounded by presence of maternal antibody), and early detection before seroconversion.

CHAPTER 5 : Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing

5.1 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 nine major proteins via alternative splicing. Our data showed that the clinical isolate $HIV_{89.6}$ produces at least 109 different spliced RNAs, including a previously unappreciated $\sim$1 kb 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.

5.2 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 [64, 187, 301–303]. 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 [65].

The integrated HIV-1 provirus is $\sim$9700 bp in length and has a single transcription

??

Figure 5.1: Mapping the splice donors and acceptors of $HIV_{89.6}$. PacBio sequence reads of $HIV_{89.6}$ cDNA from infected HOS-CD4-CCR5 (HOS) and $CD4^+$ T cells were aligned to the $HIV_{89.6}$ genome shown in (A). Exons of the conserved HIV-1 transcripts are colored according to the encoded gene. Non-coding exons 2 and 3 are variably included in each transcript where indicated. Conserved (black) and published cryptic (brown) splice donors ('D') and acceptors ('A') are shown. Numbering is according to previous convention [57]. 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 $HIV_{89.6}$ 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 for each cell type. Putative splice donors and acceptors, 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. Coordinates of the splice donors and splice acceptors are provided in Supplementary Table S2. The novel acceptor A8c was further verified. Asterisks indicate putative splice donors and acceptors that are adjacent to dinucleotides other than the consensus GT and AG, respectively.

start site, but according to the published literature yields at least 47 different mRNAs encoding 9 proteins or polyproteins, making HIV an attractive model for studies of alternative splicing [57]. HIV mRNAs fall into three classes: the unspliced RNA genome, which encodes Gag/Gag-Pol; partially spliced transcripts, $\sim$4 kb in length, encoding Vif, Vpr, a one-exon version of Tat, and Env/Vpu; and completely spliced mRNAs of roughly 2 kb encoding Tat, Rev and Nef (Figure **??**A). Additional rare 'cryptic' splice donors (5′ splice sites) and acceptors (3′ splice sites) contribute even more mRNAs [59, 113, 216, 217, 304, 305]. 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 [34, 37, 38, 43, 306–309]. Studies have suggested strain-specific splicing patterns may exist [57, 310, 311]. 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 [64, 312, 313]. Importantly for HIV, these include changes during T-cell activation [314, 315], raising the question of how cell-specific splicing affects HIV replication.

While most studies of HIV-1 splicing have been conducted in cell lines using lab-adapted viral strains, limited works in PBMCs from infected patients, monocytes and macrophages have suggested that differences may indeed exist in relevant cell types [132, 217, 310, 316]. Moreover, human splicing patterns differ between individuals, but such polymorphisms have not been investigated in the context of HIV infection [62, 63].

Here, we use deep sequencing to comprehensively characterize the transcriptome of an early passage clinical isolate, $HIV_{89.6}$ [134], in primary $CD4^+$ 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) sequencing technology, which yields read lengths up to 10 kb [74]. 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 $HIV_{89.6}$ 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.

5.3 Materials and methods

5.3.1 Cell culture and viral infections

$HIV_{89.6}$ was generated by transfection and subsequent expansion in SupT1 cells. Primary T cells were isolated by the University of Pennsylvania Center for AIDS Research Immunology core and confirmed to be homozygous for the wild-type CCR5 allele as

shown in Supplementary Table S1 and described in Supplementary Methods. HOS-CD4-CCR5 cells [317, 318] were obtained through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH from Dr Nathaniel Landau. Single round infections in T cells and HOS-CD4-CCR5 cells were performed using standard methods (see Supplementary Methods).

5.3.2 RNA and reverse transcription

Total cellular RNA was purified using the Illustra RNA kit (GE Life Sciences, Fairfield, CT, USA) from 5×10^6 cells per infection. Viral cDNA was made using a reverse transcription primer complementary to a sequence in U3 (RTprime, Supplementary Table S2). 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 the manufacturer's 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 $HIV_{89.6}$ (keo003, keo004, Supplementary Table S2, data not shown).

5.3.3 Bulk RT-PCR and cloning

Transcripts were amplified from cellular RNA using the Onestep RT-PCR kit (Qiagen) with primer pairs keo056/keo057 and keo058/keo059 (Supplementary Table S2) with the following amplification: 5 cycles of 30 s at 94°C, 12 s at 56°C, 40 s at 72°C; then 30 cycles of 30 s at 94°C, 14 s at 56°C, 40 s 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 30 s at 94°C, 30 s at 56°C and 45 s 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 (keo056/keo057 amplified material). 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 cDNAs for *tat*ˆ8c, *tat* (1 and 2 exon), *ref*, *rev* and *nef*, and the transcript with exon structure 1-5-8c were cloned into the expression vector pIRES2-AcGFP1 (Clonetech) as described in Supplementary Methods.

5.3.4 Assays of protein activity and HIV replication

Activity and HIV replication assays were performed as described in Supplementary Methods. Tat activity expressed from each cDNA was measured in TZM-bl cells [319] (gift of Dr Robert W. Doms). Rev activity was assayed in HEK-293T cells co-transfected with pCMVGagPol-RRE-R, a reporter plasmid from which Gag and Pol are expressed in a Rev-dependent manner (gift of David Rekosh) [320]. Intracellular and released supernatant p24 was measured from cells transfected with expression constructs and infected with $HIV_{89.6}$.

5.3.5 Western blotting

HEK-293T cells were transfected with expression constructs and treated with MG132 (EMD Chemicals) to inhibit the proteosome or DMSO (Supplementary Methods). 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) [321]. Horseradish peroxidase (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).

5.3.6 Single-molecule amplification

Amplification was performed by RainDance Technologies using a protocol similar to that previously reported (detailed description in Supplementary Methods) [71]. Amplification was carried out in droplets to suppress competition between amplicons. PCR droplets were generated on the RDT 1000 (RainDance Technologies) 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 (Supplementary Table S2).

5.3.7 Single-molecule sequencing

DNA amplification 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 [74]. Sequence information was acquired during real time as the immobilized DNA polymerase translocated along the template molecule. 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 [322]), allowing error control by calculating the consensus ('circular consensus sequence' or CCS). For raw reads, the average length was 2860 nt, and 10% were > 5000 nt. After condensing into consensus reads, the mean read length was 249.5 nt, due to the use of a shorter Pacific Biosciences sequencing protocol to accommodate the small size of many amplicons. Consensus reads of 1% were > 1100 nt. Sequencing data were collected in 45-min movies.

5.3.8 Data analysis

Raw reads were processed to produce CCSs. 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 [136]. Misprimed reads matching the RT primer,

reads with a CCS length shorter than 40 nt or raw length shorter than 100 nt and reads matching the human genome were discarded. Filtered reads were aligned against the $HIV_{89.6}$ 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 > 10 nt 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 nt 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 spliceform (specific transcript structure) or set of possible spliceforms. 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 spliceforms or if all other possible spliceforms contained rare ($< 1\%$ usage) donors or

acceptors (Supplementary Table S3). Otherwise, the read was called indeterminate.

To calculate the ratios of transcripts within the partially spliced class, we counted the number of reads for each assigned spliceform amplified by primer pair 1.3 and divided by the total number of assigned partially spliced reads amplified with these primers (Supplementary Figure S1 and Supplementary Table S2). 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 ($\sim$2 and $\sim$1 kb). To compare ratios of $\sim$2 kb 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 five 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 five reads and not previously assigned to a particular transcript (Figure 2) was counted as evidence of an additional transcript (47 additional junctions were detected, see Supplementary Table S4). 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, resulting in 31 additional transcripts. The minimum transcript number calculated by a greedy algorithm treating introns as events in a scheduling problem agreed with the above calculation.

Several groups have demonstrated tissue- and differentiation-specific splicing of cellular genes [64, 312, 313]. Importantly for HIV, these include changes during T-cell activation [314, 315], raising the question of how cell-specific splicing affects HIV replication. While most studies of HIV-1 splicing have been conducted in cell lines using lab-

??

Figure 5.2: Spliced transcripts produced from $HIV_{89.6}$. $HIV_{89.6}$ transcripts in T cells for which the full message structure was determined are shown arranged by size class (unspliced genome, partially spliced or 4 kb, completely spliced or 2 kb, and a new completely spliced 1 kb class). Thick bars correspond to exons and 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, three were detected in fewer than five reads (one *tat* and two *env/vpu* messages, indicated, $\diamond$) and two messages were not detected and are not shown (one encoding Vpr and one encoding Env/Vpu). Depicted non-conserved transcripts (using novel or cryptic splice sites) were each detected in at least five independent sequence reads across samples from at least two different human T-cell donors.

adapted viral strains, limited works in PBMCs from infected patients, monocytes and macrophages have suggested that differences may indeed exist in relevant cell types [132, 217, 310, 316]. Moreover, human splicing patterns differ between individuals, but such polymorphisms have not been investigated in the context of HIV infection [62, 63].

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.

5.3.9 Statistical analysis

Statistical modeling was performed using generalized linear modeling as described in Supplementary Report S2. All analyses were performed in R 2.14.0 (R Development Core) [139].

5.3.10 Data access

Sequence data is available in the SRA database with the following accession numbers: SRP014319.

5.4 Results

5.4.1 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 $CD4^+$ T cells of seven different healthy human donors infected in vitro with $HIV_{89.6}$, an early passage dual-tropic clade-B clinical isolate (Supplementary Figure S1, human donor data in Supplementary Table S1) [134]. 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 h infected T cells.

To preserve the relative proportions of template molecules while amplifying the cDNA, we used RainDance Technologies' single-molecule micro-droplet based PCR [71]. Droplet libraries containing multiple overlapping primer pairs were designed to query all message forms and allow later calculation of relative abundance (Supplementary Table S2 and Supplementary Figure S1). 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 [74, 322]. 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.

5.4.2 Splice donors and acceptors

We aligned PacBio reads containing HIV sequences to the $HIV_{89.6}$ 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 (Figure **??**A, 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 (Figure **??** and Supplementary Table S3). In order to be selected as a bona fide splice site and remove artifacts possibly created by recombination during sample preparation, 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. The most frequently used novel splice site was an acceptor that we have termed A8c because it lies near A8, A8a and A8b (discussed in detail below). Additional novel sites are further discussed in Supplementary Report S1.

Most of the new splice sites adhered to consensus sequences for the standard spliceosome (Supplementary Table S3). However, there appeared to be one splice donor upstream of D1 with a cytidine in place of the usual uracil 2 nt downstream of the splice site. Similar 'GC donors' appear in 1% of known splice junctions in humans [323]. Of the novel splice acceptors, three were preceded by dinucleotides other than the consensus AG. Alternative dinucleotides are used infrequently as splice acceptors [324–327]; however, it is possible that our deep sequencing method allowed us to observe rare events.

5.4.3 Structures of spliced $HIV_{89.6}$ RNAs

To quantify the populations of HIV-1 transcripts, we aligned all reads to the collection of 47 well-established spliced HIV-1 transcripts and detected 45 of them (Figure 2). We additionally aligned reads to the $HIV_{89.6}$ genome allowing all possible combinations of splice junctions—canonical, cryptic or novel—determined from the sequencing data (Figure **??**), yielding an additional 32 complete transcripts, 19 of which were novel. The data also provide evidence for more novel splice junctions but in incomplete sequences,

implying the existence of additional new transcripts (Supplementary Table S4 and Supplementary Report S1). The full data set taken together provides evidence for least 109 different $HIV_{89.6}$ transcripts in primary T cells.

Amplification primers that isolated the two main classes of spliced messages allowed us to determine the ratios of mRNAs in each (Figure 2 and Supplementary Table S5). 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 $\sim$2 kb 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 $\sim$2 kb messages (Figure 2) [57, 328]. 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. Due to biases inherent in the reverse transcription step, we could only compare transcripts within each size class, and we note that our methods have not been validated for empirical quantification. However, the ratios were roughly confirmed using overlapping sequence reads obtained with alternate primer pairs and by end point RT-PCR analysis of HIV-1 RNAs (data not shown).

Exons 2 and 3 are non-coding exons whose inclusion in transcripts other than *vif* and *vpr* has no known function. We found that they were included in other messages infrequently, each in $\sim$7–8% of transcripts in the $\sim$2 kb completely spliced class of transcripts and 5% of partially spliced transcripts accumulating in T cells. This is consistent with previous measurements in the partially spliced class but much lower than has been estimated for completely spliced transcripts in HeLa cells, suggesting cell-type-specific splicing patterns may influence inclusion of these exons [57].

5.4.4 A novel ~ 1 kb class of completely spliced transcripts

Primers placed near the 5′- and 3′-ends of the $\text{HIV}_{89.6}$ genome amplified a second class of completely spliced transcripts ~ 1 kb in length. In place of A7, these transcripts use a set of little studied splice acceptors located ~ 800 bp downstream within the 3′-TR. Two groups have previously observed splicing from D1 to acceptors A8, A8a and A8b in this region, yielding messages of this size class in patient samples; however, none of these could be translated to a protein of significant length [216, 217]. We determined the complete structure of 29 members of the 1-kb class (Figure 2 and Supplementary Table S5). The most abundant messages observed in this class use the novel acceptor A8c to define their terminal exon. For HIV89.6, acceptor A8c was used nearly as frequently as A7, which gives us the 2-kb class of transcripts (Supplementary Table S3), and this was supported by end point 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%) [329]. This is due to the poor conservation of an adenosine at the wobble base position of the 123rd codon (proline) of the Nef reading frame, which creates the AG dinucleotide generally required at splice acceptors. Since any base at this position would code for proline, there does not seem to be strong selection for a splice acceptor here. However, A8c is displaced from nearby well-conserved ($> 90\%$) cryptic acceptors A8a and A8b by multiples of 3 bp (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 [329]. We confirmed that the 1 kb transcripts using A8a, A8b and A8c were present in infected HOS and T cells by end point RT-PCR using additional primer pairs and by Sanger sequencing of cloned transcripts (Figure 3A and B; 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

??

Figure 5.3: $HIV_{89.6}$ transcripts were amplified by RT-PCR using RNA from infected HOS-CD4-CCR5 cells with primers keo056 and keo057 (Supplementary Table S2). Major bands detected after gel electrophoresis were cloned from the 48 hpi sample and message structures determined by Sanger sequencing. Thick bars represent exons and dashed lines excised introns. Genes are shown above (not to scale) with start codons indicated by circles. Coding potentials of open reading frames are described. The first two start codons in messages 5 and 6, circles below, are not shared by known HIV-1 genes. Messages 1, 2, 4 and 5 were cloned into expression plasmids for activity assays. (B) Confirmation of presence of the $\sim$1 kb message RNAs in HOS-CD4-CCR5 and primary $CD4^+$ T cells (human donor 1, harvested 24 and 48 hpi). An independent primer pair (keo058 and keo059) was used to amplify transcripts by RT-PCR. Expected amplicon sizes for transcripts in (A) are shown. (C) Tat activity was measured in Tzm-bl cells as Tat-dependent luciferase production after transient transfection with expression plasmids. (D) Western blot showing expression of protein of the predicted size for Ref (12.5 kb) in cells transfected with the Ref expression construct and treated with proteosome inhibitor MG132, detected by an antibody recognizing the carboxy-terminus of Nef. Expression plasmid encoding Nef was included to control for possible expression of partial Nef peptides or breakdown products from the Nef ORF.

novel amino acids (termed Tatˆ8c). Tatˆ8c showed activity when overexpressed in cells containing a Tat reporter construct (Figure 3C, nucleotide and amino acid sequences in Supplementary Table S6). Transcripts with exon structures 1-4a/b/c-8c encode a novel fusion of the amino-terminal 26 amino acids of Rev and the carboxy-terminal 80 amino acids of Nef, hereafter referred to as Ref. 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 (Supplementary Figure S2). 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 (Figure 3D). 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.

5.4.5 Temporal dynamics of transcript populations

To assess longitudinal variation, we investigated $HIV_{89.6}$ transcript populations during the course of a single round of infection in HOS-CD4-CCR5 cells. A sensitive method for

??

Figure 5.4: Temporal, cell type and donor variability in accumulation of HIV-1 messages. (A) In order to highlight changes in ratios of HIV-1 transcripts accumulating over time during infection and between HOS-CD4-CCR5 cells and primary T cells, we used PacBio read counts to calculate proportions of transcripts with splicing from the first major splice donor, D1, to each of the mutually exclusive acceptors: A3 (required to make Tat), A4c, A4a, A4c (Env/Vpu and Rev), A5 (Env/Vpu and Nef) and the novel putative acceptor A5a. Sequences used in the analysis derived from templates amplified with primers F1.2 and R1.2 (Supplementary Table S2). The heat map shows average data for T cell and HOS cell samples in columns with the color tiles indicating the proportion of D1 splicing to each of the mutually exclusive acceptors (rows), according to the color scale shown. Statistics for this analysis based on a generalized linear model are provided in Supplementary Report S2. (B) Reverse transcription and bulk PCR amplification of $HIV_{89.6}$ transcripts from HOS cells and primary T cells from one human subject (subject 3) resolved by agarose gel electrophoresis and stained with ethidium bromide verified temporal and cell type changes shown in (A).

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 h that were statistically significant ($p < 10^{-10}$; generalized linear model described in Supplementary Report S2). This pattern was especially evident in junctions involving donor 1 spliced to each of these acceptors (Figure 4A). Most dramatically, transcripts with splicing junctions between D1 and A3 (tat messages) increased with time ($p < 10^{-10}$), while D1ˆA4b 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 end point RT-PCR and analysis after electrophoresis on ethidium-stained gels (Figure 4B).

5.4.6 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 D1ˆA5 junctions were dominant in both cell types, messages using the D1ˆA4c 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 (Figure 4A). 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 (Figure 4B).

5.4.7 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 (Figure 4A). The magnitudes of predicted differences were small, all $< 33\%$ and most $< 10\%$.

5.5 Discussion

Use of single-molecule enrichment and long-read single-molecule sequencing has 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 HIV89.6 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 population. 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. 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 ($\sim$1 kb) 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. $\text{HIV}_{89.6}$ is a particularly cytotoxic virus isolated from the CSF of a patient, and it forms unusually large syncitia in macrophages [134]. The abundance of 1-kb 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 require further studied to clarify possible functions.

The wealth of new messages found here in $\text{HIV}_{89.6}$ 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 [34], mutations creating sequences that even slightly resemble the 3′ splice site consensus may be occasionally recruited as novel acceptors, creating new mRNAs. 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 bp, respectively [330], though sequence substitutions in HIV are usually also constrained by overlapping viral coding regions. We and others have observed appearance of novel exons within the major HIV-1 introns [59, 304, 305]. 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. Clusters of splice sites might also provide redundancies that protect vital messages, as suggested previously [331, 332]. Frequent evolution of new splice sites may allow viruses to test out new combinations of exons, potentially yielding new RNAs and 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.

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 [43, 307]. 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 [93, 333]. Inhibition of SF2/ASF has already been shown to abrogate HIV-1 replication in vitro [43]. Thus the lability seen here for function of these factors suggests they may be attractive antiretroviral targets.

5.6 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 ex-

pertise. F.D.B., K.T., D.L., E.S., K.E.O. and R.M. conceived and designed the experiment. K.E.O. and R.C.A. carried out sample preparation and experimental validation. P.D. and J.O. performed single-molecule amplification. K.T. and S.W. performed sequencing. S.S.-M., K.E.O. and M.B. analyzed the data. K.E.O., F.D.B. and S.S.-M. wrote the manuscript.

CHAPTER 6 : Conclusions and future directions

PacBio was bad. Figure? Do better

Comparison among labs and cell types/viruses at same time. Standardize.

Non polyadenylated RNA. Strand specific sequencing. Longer reads and longer fragments.

Ebola deaths and infected. Ebola LAMP figure 6.1.

In addition an important subset of HIV are the founder viruses transmitted between hosts [334, 335]. These viruses are not well studied and perhaps their splicing and gene expression differ from the rest of the viral swarm of late-term patients.

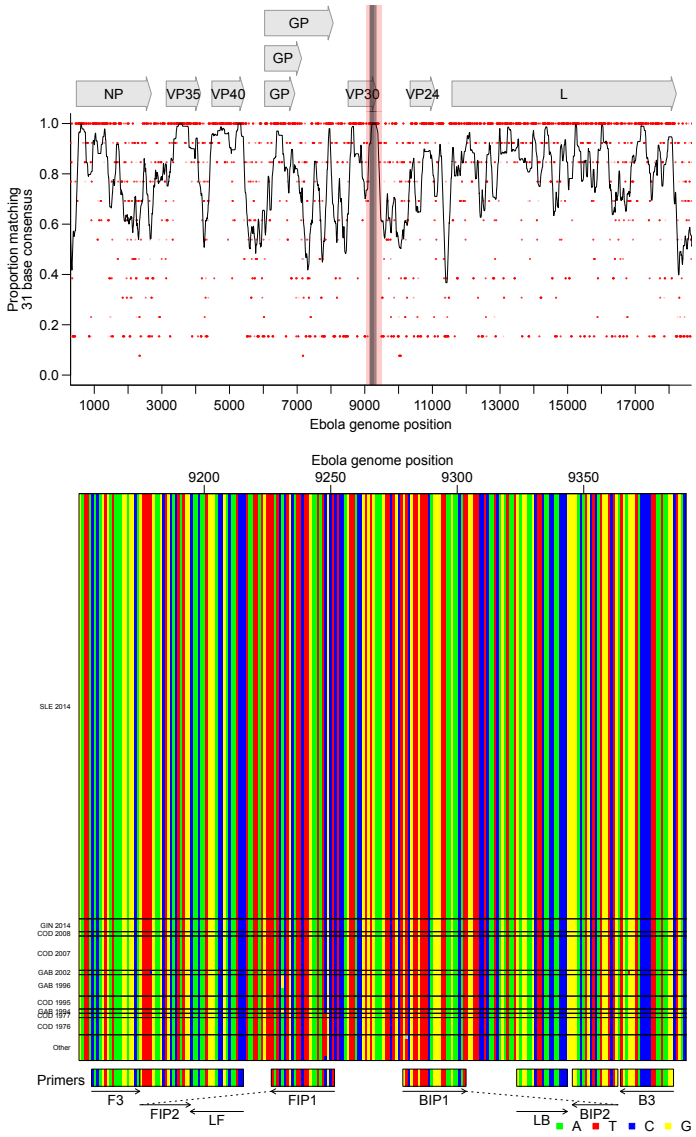


Figure 6.1: Bioinformatic analysis to design Ebola RT-LAMP primers. A) Conservation of sequence in Ebola. Ebola genomes (n = [[number]]) from the [[XXX]] collection were aligned and conservation calculated. The x-axis shows the coordinate on the Ebola genome, the y-axis shows the proportion of sequences matching the consensus for each 21 base segment of the genome (red points). The black line shows a 101 base sliding average over these proportions. The vertical red shading shows the region targeted for LAMP primer design that was used as input into the EIKEN primer design tool. Numbering is relative to the [[XX]] sequence. B) Aligned genomes, showing the locations of the preliminary primers. Sequences in the red shaded region in A are shown, with DNA bases color-coded as shown at the lower right. Each row indicates an HIV sequence and each column a base in that sequence. Horizontal lines separate the HIV subtypes (labeled at right). Arrows indicate the strand targeted by each primer. Primers targeting the negative strand of the virus are shown as reverse compliments for ease of viewing.

APPENDIX A.1 : Reproducible report of HIV integration sites and latency analysis

A.1.1 Supplementary data

Additional File 2 is a gzipped csv file that includes a row for each uniquely mapped provirus and its surrounding genomic annotations. The csv file should have 12436 rows (excluding header) with 6252 expressed and 6184 latent proviruses.

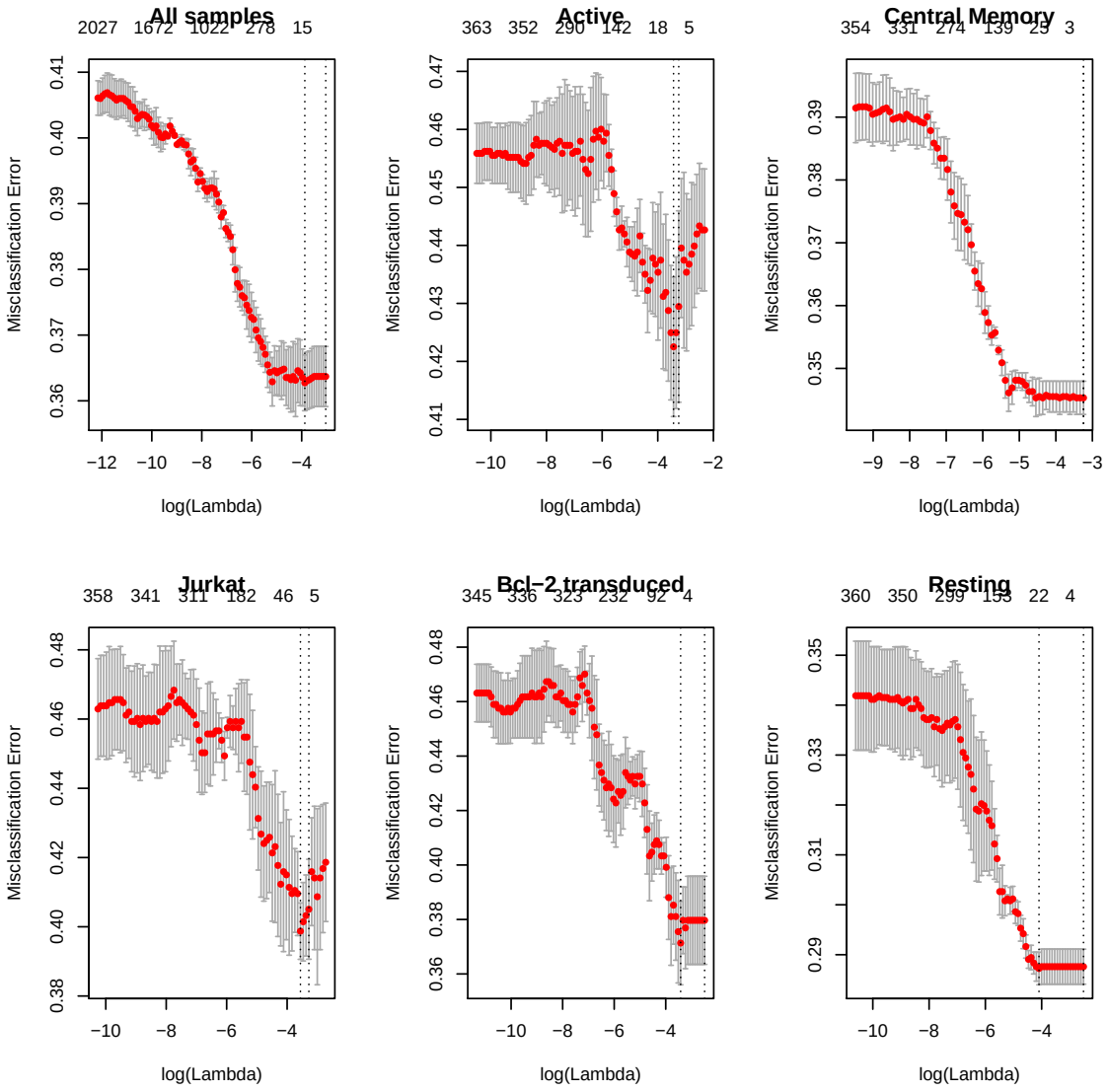
```
integrationData <- read.csv("AdditionalFile2.csv.gz",
   stringsAsFactors = FALSE)

nrow(integrationData)

## [1] 12436

table(integrationData$isLatent)

##
## FALSE  TRUE
##  6252  6184
```

A.1.2 Lasso regression

The lasso regressions take a while to run so I've turned down the number of cross validations here (set `eval=FALSE` below to completely skip this step). Leave one out and 480-fold cross validation were used in the paper but processing may take a few days without parallel processing. Lasso regression requires the R glmnet package.

```
notFitColumns <- c("id", "chr", "pos", "strand", "sample", "
   isLatent")

samples <- unique(as.character(integrationData$sample))

sampleMatrix <- do.call(cbind, lapply(samples, function(x)
   integrationData$sample ==
    x))

colnames(sampleMatrix) <- gsub(" ", "_", samples)

interact <- function(predMatrix, columns, addNames = NULL) {
    out <- do.call(cbind, lapply(1:ncol(columns), function(x)
       predMatrix *
        columns[, x]))
    if (!is.null(addNames)) {
        if (length(addNames) != ncol(columns)) {
            stop(simpleError("Names not same length as columns"))
        }
        colnames(out) <- sprintf("%s_%s", rep(addNames, each =
           ncol(predMatrix)),
            rep(colnames(predMatrix), length(addNames)))
    }
    return(out)
}

fitData <- as.matrix(integrationData[, !colnames(integrationData)
    %in%
    notFitColumns])
```

```
library(glmnet)

penalties <- rep(1, ncol(fitData2))

penalties[ncol(fitData2) - (ncol(sampleMatrix):1) + 1] <- 0

lassoFit <- cv.glmnet(fitData2, integrationData$isLatent, family
   = "binomial",
    type.measure = "class", nfolds = 3, penalty.factor =
       penalties)

seperateFits <- lapply(samples, function(x) cv.glmnet(fitData[
   integrationData$sample ==
    x, ], integrationData$isLatent[integrationData$sample ==
    x], family = "binomial", type.measure = "class", nfolds = 3))

names(seperateFits) <- samples
```



A.1.3 Correlation

We looked for correlation between the genomic variables and expression status of the proviruses.

```
corMat <- apply(fitData, 2, function(x) sapply(samples, function(
   y) {
    selector <- integrationData$sample == y
```

```
    if (sd(x[selector]) == 0)
        return(0)
    isLatent <- integrationData[selector, "isLatent"]
    cor(as.numeric(isLatent), x[selector], method = "spearman")
}))

quantile(corMat, seq(0, 1, 0.1))

##           0%          10%          20%          30%
## -0.185223020 -0.081555830 -0.048938130 -0.030895834
##          40%          50%          60%          70%
## -0.018053321 -0.005613895  0.003580982  0.017822483
##          80%          90%         100%
##  0.036694554  0.062003356  0.170642314
```

If we looked for genomic variables consistently correlated or anti-correlated with proviral expression status with an FDR q-value less than 0.01, no variable was significantly correlated in more than 3 samples.

```
pMat <- apply(fitData, 2, function(x) sapply(samples, function(y)
    {
    selector <- integrationData$sample == y
    if (sd(x[selector]) == 0)
        return(NA)
    isLatent <- integrationData[selector, "isLatent"]
    cor.test(as.numeric(isLatent), x[selector], method = "
       spearman",
        exact = FALSE)$p.value
```

```
}))

adjustPMat <- pMat

adjustPMat[, ] <- p.adjust(pMat, "fdr")

downPMat <- upPMat <- adjustPMat

downPMat[corMat > 0] <- 1

upPMat[corMat < 0] <- 1

table(apply(upPMat < 0.01 & !is.na(upPMat), 2, sum))

##
##   0   1   2   3
## 298  27  38  10

table(apply(downPMat < 0.01 & !is.na(downPMat), 2, sum))

##
##   0   1   2   3
## 216  36  63  58
```

A.1.4 RNA expression

We fit a logistic regression to a polynomial of log RNA-Seq reads within 5000 bases from Jurkat cells for the Jurkat sample and T cells for the rest.

```
rna <- ifelse(integrationData$sample == "Jurkat",
   integrationData$log_jurkatRNA,
    integrationData$rna_5000)
```

```
rna2 <- rna^2

rna3 <- rna^3  #

rna4 <- rna^4

glmData <- data.frame(isLatent = integrationData$isLatent, sample
    = integrationData$sample,
    rna, rna2, rna3, rna4)

glmMod <- glm(isLatent ~ sample * rna + sample * rna2 + sample *
    rna3 + sample * rna4, data = glmData, family = "binomial")

summary(glmMod)

##
## Call:
## glm(formula = isLatent ~ sample * rna + sample * rna2 + sample
    *
##     rna3 + sample * rna4, family = "binomial", data = glmData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2899  -0.9864  -0.8676   1.0960   1.6007
##
## Coefficients:
##                               Estimate Std. Error z value
## (Intercept)                  1.7623655  0.2138859   8.240
## sampleBcl-2 transduced      -2.1625912  0.7061524  -3.062
## sampleCentral Memory        -2.5010063  0.2437685 -10.260
## sampleJurkat                -2.0800202  0.2836871  -7.332
```

```
## sampleResting                0.7840481  0.3312247   2.367
## rna                         -0.6567268  0.2344422  -2.801
## rna2                         0.1387703  0.0770589   1.801
## rna3                        -0.0167219  0.0094076  -1.777
## rna4                         0.0007572  0.0003845   1.969
## sampleBcl-2 transduced:rna   0.5750186  0.6366537   0.903
## sampleCentral Memory:rna     0.9067758  0.2750955   3.296
## sampleJurkat:rna             0.5294036  0.3867163   1.369
## sampleResting:rna            0.0366276  0.3436248   0.107
## sampleBcl-2 transduced:rna2 -0.0369353  0.1878816  -0.197
## sampleCentral Memory:rna2   -0.2106715  0.0915492  -2.301
## sampleJurkat:rna2           -0.0766215  0.1641153  -0.467
## sampleResting:rna2          -0.0760450  0.1086998  -0.700
## sampleBcl-2 transduced:rna3  0.0032503  0.0213743   0.152
## sampleCentral Memory:rna3    0.0237064  0.0112661   2.104
## sampleJurkat:rna3            0.0042183  0.0263910   0.160
## sampleResting:rna3           0.0153132  0.0128711   1.190
## sampleBcl-2 transduced:rna4 -0.0002532  0.0008267  -0.306
## sampleCentral Memory:rna4   -0.0009877  0.0004627  -2.135
## sampleJurkat:rna4            0.0001725  0.0014215   0.121
## sampleResting:rna4          -0.0008049  0.0005119  -1.572
##                            Pr(>|z|)
## (Intercept)                 < 2e-16 ***
## sampleBcl-2 transduced       0.00219 **
## sampleCentral Memory        < 2e-16 ***
## sampleJurkat                2.27e-13 ***
## sampleResting                0.01793 *
```

```
## rna                           0.00509 **
## rna2                          0.07173 .
## rna3                          0.07549 .
## rna4                          0.04891 *
## sampleBcl-2 transduced:rna    0.36643
## sampleCentral Memory:rna      0.00098 ***
## sampleJurkat:rna              0.17101
## sampleResting:rna             0.91511
## sampleBcl-2 transduced:rna2   0.84415
## sampleCentral Memory:rna2     0.02138 *
## sampleJurkat:rna2             0.64059
## sampleResting:rna2            0.48419
## sampleBcl-2 transduced:rna3   0.87913
## sampleCentral Memory:rna3     0.03536 *
## sampleJurkat:rna3             0.87301
## sampleResting:rna3            0.23415
## sampleBcl-2 transduced:rna4   0.75939
## sampleCentral Memory:rna4     0.03280 *
## sampleJurkat:rna4             0.90339
## sampleResting:rna4            0.11585
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17240  on 12435  degrees of freedom
```

```
## Residual deviance: 15874  on 12411  degrees of freedom
## AIC: 15924
##
## Number of Fisher Scoring iterations: 4
```

A.1.5 Strand orientation

We used a Fisher's exact test to check if silent/inducible proviruses were enriched when integrated in the same strand orientation as cellular genes.

```
selector <- integrationData$inGene == 1

strandTable <- with(integrationData[selector, ], table(ifelse(
   isLatent,
    "Silent/Inducible", "Active"), ifelse(inGeneSameStrand ==
    1, "Same", "Diff"), sample))

apply(strandTable, 3, fisher.test)

## $Active
##
##      Fisher's Exact Test for Count Data
##
## data:  array(newX[, i], d.call, dn.call)
## p-value = 0.06061
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.7219466 1.0081995
## sample estimates:
## odds ratio
```

```
##  0.8532127
##
##
## $`Bcl-2 transduced`
##
##      Fisher's Exact Test for Count Data
##
## data:  array(newX[, i], d.call, dn.call)
## p-value = 2.177e-05
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.446896 2.872562
## sample estimates:
## odds ratio
##   2.036148
##
##
## $`Central Memory`
##
##      Fisher's Exact Test for Count Data
##
## data:  array(newX[, i], d.call, dn.call)
## p-value = 0.2907
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.9386167 1.2320238
## sample estimates:
```

```
## odds ratio
##    1.07529
##
##
## $Jurkat
##
##      Fisher's Exact Test for Count Data
##
## data:  array(newX[, i], d.call, dn.call)
## p-value = 0.1674
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.9207548 1.5699893
## sample estimates:
## odds ratio
##   1.202007
##
##
## $Resting
##
##      Fisher's Exact Test for Count Data
##
## data:  array(newX[, i], d.call, dn.call)
## p-value = 0.5732
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.7825231 1.1405158
```

```
## sample estimates:
## odds ratio
##  0.9447415
```

A.1.6 Acetylation

To reduce correlation between acetylation marks, we generated the first ten principal components of the acetylation data and ran a logistic regression against them. We compared the cross validated performance of this regression with a base model only including which dataset the integration site came from. The cross-validation here has been reduced for efficiency but 480-fold cross-validation was used in the paper.

```
acetyl <- integrationData[, !grepl("logDist", colnames(
   integrationData)) &
    grepl("ac", colnames(integrationData))]

acetylPCA <- princomp(acetyl)

cumsum(acetylPCA$sdev[1:10]^2/sum(acetylPCA$sdev^2))

##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
## 0.5947268 0.6786611 0.7267433 0.7610502 0.7833616 0.7964470
##    Comp.7    Comp.8    Comp.9   Comp.10
## 0.8093295 0.8215027 0.8299358 0.8372584

cv.glm <- function(model, K = nrow(thisData), subsets = NULL) {
    modelCall <- model$call
    thisData <- eval(modelCall$data)
    n <- nrow(thisData)
    if (is.null(subsets))
```

```
        subsets <- split(1:n, sample(rep(1:K, length.out = n)))
    preds <- lapply(subsets, function(outGroup) {
        subsetData <- thisData[-outGroup, , drop = FALSE]
        predData <- thisData[outGroup, , drop = FALSE]
        thisModel <- modelCall
        thisModel$data <- subsetData
        return(predict(eval(thisModel), predData))
    })
    pred <- unlist(preds)[order(unlist(subsets))]
    subsetId <- rep(1:K, sapply(subsets, length))[order(unlist(
       subsets))]
    return(data.frame(pred, subsetId))
}

inData <- data.frame(isLatent = integrationData$isLatent, sample
   = as.factor(integrationData$sample),
    acetylPCA$score[, 1:10])

modelPreds <- cv.glm(glm(isLatent ~ sample + Comp.1 + Comp.2 +
    Comp.3 + Comp.4 + Comp.5 + Comp.6 + Comp.7 + Comp.8 + Comp.9
       +
    Comp.10, family = "binomial", data = inData), K = 5)

basePreds <- cv.glm(glm(isLatent ~ sample, family = "binomial",
    data = inData), subsets = split(1:nrow(inData),
       modelPreds$subsetId),
    K = 5)

modelCorrect <- sum((modelPreds$pred > 0) ==
   integrationData$isLatent)
```

```
baseCorrect <- sum((basePreds$pred > 0) ==
   integrationData$isLatent)

prop.test(c(baseCorrect, modelCorrect), rep(nrow(integrationData)
   ,
    2))

##
##      2-sample test for equality of proportions with
##      continuity correction
##
## data:  c(baseCorrect, modelCorrect) out of rep(nrow(
   integrationData), 2)
## X-squared = 0.00017372, df = 1, p-value = 0.9895
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01187726  0.01219890
## sample estimates:
##    prop 1    prop 2
## 0.6362978 0.6361370
```

A.1.7 Gene deserts

We used Fisher's exact test to look for an association between integration outside a gene and proviral expression status.

```
geneTable <- table(integrationData$isLatent,
   integrationData$inGene,
    integrationData$sample)
```

```
apply(geneTable, 3, fisher.test)

## $Active
##
##      Fisher's Exact Test for Count Data
##
## data:  array(newX[, i], d.call, dn.call)
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.3629548 0.5446204
## sample estimates:
## odds ratio
##  0.4452621
##
##
## $`Bcl-2 transduced`
##
##      Fisher's Exact Test for Count Data
##
## data:  array(newX[, i], d.call, dn.call)
## p-value = 0.1052
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.9203418 2.3478599
## sample estimates:
## odds ratio
##   1.472224
```

```
##
##
## $`Central Memory`
##
##      Fisher's Exact Test for Count Data
##
## data:  array(newX[, i], d.call, dn.call)
## p-value = 0.7803
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.8525329 1.1253952
## sample estimates:
## odds ratio
##  0.9791165
##
##
## $Jurkat
##
##      Fisher's Exact Test for Count Data
##
## data:  array(newX[, i], d.call, dn.call)
## p-value = 0.5443
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.7909269 1.6167285
## sample estimates:
## odds ratio
```

```
##   1.127836
##
##
## $Resting
##
##      Fisher's Exact Test for Count Data
##
## data:  array(newX[, i], d.call, dn.call)
## p-value = 3.071e-08
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4384828 0.6864112
## sample estimates:
## odds ratio
##  0.5500205
```

We used a two-sample t-test to investigate whether there was a significant difference in distance to the nearest gene between expressed and silent/inducible proviruses integrated outside genes.

```
geneDistData <- integrationData[!integrationData$inGene, c("
   isLatent",
    "logDist_nearest", "sample")]

by(geneDistData, geneDistData$sample, function(x) t.test(
   logDist_nearest ~
    isLatent, data = x))

## geneDistData$sample: Active
```

```
##
##      Welch Two Sample t-test
##
## data:  logDist_nearest by isLatent
## t = -2.4539, df = 287.73, p-value = 0.01472
## alternative hypothesis: true difference in means is not equal
   to 0
## 95 percent confidence interval:
##  -0.80738340 -0.08867607
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            9.608737           10.056767
##
## ------------------------------------------------
## geneDistData$sample: Bcl-2 transduced
##
##      Welch Two Sample t-test
##
## data:  logDist_nearest by isLatent
## t = 0.40978, df = 86.2, p-value = 0.683
## alternative hypothesis: true difference in means is not equal
   to 0
## 95 percent confidence interval:
##  -0.6309351  0.9586004
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            9.036872            8.873039
```

```
##
## ------------------------------------------------
## geneDistData$sample: Central Memory
##
##      Welch Two Sample t-test
##
## data:  logDist_nearest by isLatent
## t = -0.07188, df = 861.61, p-value = 0.9427
## alternative hypothesis: true difference in means is not equal
   to 0
## 95 percent confidence interval:
##  -0.2371374  0.2203819
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            10.19225            10.20063
##
## ------------------------------------------------
## geneDistData$sample: Jurkat
##
##      Welch Two Sample t-test
##
## data:  logDist_nearest by isLatent
## t = -1.8217, df = 139.56, p-value = 0.07064
## alternative hypothesis: true difference in means is not equal
   to 0
## 95 percent confidence interval:
##  -1.26342086  0.05167979
```

```
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            9.925782           10.531652
##
## ------------------------------------------
## geneDistData$sample: Resting
##
##      Welch Two Sample t-test
##
## data:  logDist_nearest by isLatent
## t = -5.1275, df = 193.49, p-value = 7.096e-07
## alternative hypothesis: true difference in means is not equal
   to 0
## 95 percent confidence interval:
##  -1.2687917 -0.5638568
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            9.489931           10.406255
```

To check for a relationship between silent/inducible status and distance to CpG islands, we used a two sample t-test on the logged distance and saw a significant difference between silent/inducible and expressed proviruses (before accounting for a correlation between being near CpG islands and in genes)

```
t.test(integrationData$logDist_cpg ~ integrationData$isLatent)

##
##      Welch Two Sample t-test
```

```
##
## data:  integrationData$logDist_cpg by integrationData$isLatent
## t = -2.0233, df = 12381, p-value = 0.04306
## alternative hypothesis: true difference in means is not equal
   to 0
## 95 percent confidence interval:
##  -0.105657514 -0.001675563
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            10.16362            10.21728

sapply(unique(integrationData$sample), function(x) with(
   integrationData[integrationData$sample ==
    x, ], p.adjust(t.test(logDist_cpg ~ isLatent)$p.value, method
        = "bonferroni",
    n = 5)))

##           Active   Central Memory           Jurkat
##      0.512040457      1.000000000      1.000000000
## Bcl-2 transduced          Resting
##      1.000000000      0.005866539
```

Many CpG islands are found near genes. To account for this relationship, we used an ANOVA test including whether the integration site was inside a gene prior to including CpG islands. After including integration inside genes, CpG islands were not significantly associated with silent/inducible status of the proviruses with all samples grouped or individually after Bonferonni correction for multiple comparisons.

```
anova(with(integrationData, glm(isLatent ~ I(logDist_nearest ==
    0) + logDist_cpg, family = "binomial")), test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: isLatent
##
## Terms added sequentially (first to last)
##
##
##                          Df Deviance Resid. Df Resid. Dev
## NULL                                     12435      17240
## I(logDist_nearest == 0)  1  26.2682     12434      17213
## logDist_cpg              1   1.1328     12433      17212
##                           Pr(>Chi)
## NULL
## I(logDist_nearest == 0) 2.971e-07 ***
## logDist_cpg                0.2872
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sapply(unique(integrationData$sample), function(x) {
    p.adjust(anova(with(integrationData[integrationData$sample ==
        x, ], glm(isLatent ~ I(logDist_nearest == 0) +
           logDist_cpg,
        family = "binomial")), test = "Chisq")["logDist_cpg",
```

```
        "Pr(>Chi)"], method = "bonferroni", n = 5)
})

##           Active   Central Memory           Jurkat
##        1.0000000        1.0000000        1.0000000
## Bcl-2 transduced          Resting
##        1.0000000        0.2007788
```

A.1.8 Alphoid repeats

When analyzing repetitive elements, we treated each read as an independent observation and included reads with multiple alignments to the genome. Additional File 3 is a gzipped csv file containing a row for each read with multiple alignments and one row for each dereplicated integration site with a single alignment with the count variable indicating the number of reads dereplicated to that integration site. There should be 26,190 rows (excluding header) with 14,494 rows of expressed provirus and 11,696 rows of silent/inducible provirus.

```
repeats <- read.csv("AdditionalFile3.csv.gz", check.names = FALSE
   ,
    stringsAsFactors = FALSE)

nrow(repeats)

## [1] 26190

summary(repeats$isLatent)

##    Mode   FALSE    TRUE    NA's
## logical   14494   11696       0
```

```
notRepeatColumns <- c("id", "isLatent", "sample", "count")
```

To analyze whether there was an association between proviral expression status and integration within alphoid repeats, we used Fisher's exact test with a Bonferroni correction for five samples. For comparison, we looked at the association between proviral expression and the other repeats in the RepeatMasker database. We did not Bonferroni correct for the multiple repeat types so that the repeats could be compared with the analysis of alphoid repeats (for which we had an a priori hypothesis for an association with latency).

```
dummyX <- rep(c(TRUE, FALSE), 2)

dummyY <- rep(c(TRUE, FALSE), each = 2)

repeatData <- repeats[, !colnames(repeats) %in% notRepeatColumns]

repeatData <- repeatData[, apply(repeatData, 2, sum) > 0]

testRepeats <- function(x, repeats) {
    sapply(samples, function(thisSample, repeats) {
        selector <- repeats$sample == thisSample
        repLatent <- rep(repeats$isLatent[selector],
           repeats$count[selector])
        repRepeat <- rep(x[selector], repeats$count[selector])
        fisher.test(table(c(dummyX, repLatent), c(dummyY,
           repRepeat)) -
            1)$p.value
    }, repeats)
}
```

```
repeatPs <- apply(repeatData, 2, testRepeats, repeats[,
   notRepeatColumns])

table(apply(repeatPs * 5 < 0.05, 2, sum))

##
##   0   1   2   3
## 611  76  15   1

which(apply(repeatPs * 5 < 0.05, 2, sum) >= 3)

## ALR/Alpha
##       178

p.adjust(repeatPs[, "ALR/Alpha"], "bonferroni")

##           Active   Central Memory           Jurkat
##     5.026890e-02     3.940207e-03     1.027189e-08
## Bcl-2 transduced          Resting
##     1.000000e+00     2.424896e-02
```

A.1.9 Neighbors

We looked at all pairs of viruses on the same chromosome separated by no more than a given distance, e.g. 100 bases, either with all samples pooled or split between within sample pairs or between sample pairs.

```
allNeighbors <- data.frame(id1 = 0, id2 = 0)[0, ]

ids <- 1:nrow(integrationData)

for (chr in unique(integrationData$chr)) {
```

```
    chrSelector <- integrationData$chr == chr
    neighborPairs <- data.frame(id1 = rep(ids[chrSelector], sum(
       chrSelector)),
        id2 = rep(ids[chrSelector], each = sum(chrSelector)))
    neighborPairs <- neighborPairs[neighborPairs$id1 <
       neighborPairs$id2,
        ]
    allNeighbors <- rbind(allNeighbors, neighborPairs)
}

allNeighbors$dist <- abs(integrationData$pos[allNeighbors$id1] -
    integrationData$pos[allNeighbors$id2])

allNeighbors$latent1 <- integrationData$isLatent[allNeighbors$id1
   ]

allNeighbors$latent2 <- integrationData$isLatent[allNeighbors$id2
   ]

allNeighbors$sample1 <- integrationData$sample[allNeighbors$id1]

allNeighbors$sample2 <- integrationData$sample[allNeighbors$id2]

allNeighbors <- allNeighbors[allNeighbors$dist <= 1e+06, ]
```

The expected number of matching pairs was calculated as $\sum_{j \in \text{samples}} n_{j,d}(\theta_{j,d}\theta_{\neg j,d} + (1 - \theta_{j,d})(1 - \theta_{\neg j,d}))$ for between sample, $\sum_{j \in \text{samples}} n_{j,d}(\theta_{j,d}^2 + (1 - \theta_{j,d})^2)$ for within sample and $n_d(\theta_d^2 + (1 - \theta_d)^2)$ for all pairs, where $n_{j,d}$ is the number of pairs of proviruses separated by no more than d base pairs where the first provirus is from sample j, $\theta_{j,d}$ is the proportion of silent/inducible proviruses in sample j appearing in at least one

pair of proviruses separated by less than d base pairs and $\neg j$ means all samples except sample j.

```
dists <- unique(round(10^seq(1, 6, 1)))

pairings <- do.call(rbind, lapply(dists, function(x, allNeighbors
   ) {
    inSelector <- allNeighbors$dist <= x & allNeighbors$sample1
       ==
        allNeighbors$sample2
    outSelector <- allNeighbors$dist <= x & allNeighbors$sample1
       !=
        allNeighbors$sample2
    allSelector <- allNeighbors$dist <= x
    out <- data.frame(dist = x, observedIn = sum(allNeighbors[
       inSelector,
        "latent1"] == allNeighbors[inSelector, "latent2"]),
           observedOut = sum(allNeighbors[outSelector,
        "latent1"] == allNeighbors[outSelector, "latent2"]),
        observedAll = sum(allNeighbors[allSelector, "latent1"] ==
            allNeighbors[allSelector, "latent2"]), totalIn = sum(
               inSelector),
        totalOut = sum(outSelector), totalAll = sum(allSelector))
    out$expectedIn <- sum(with(allNeighbors[inSelector, ], sapply
       (samples,
        function(x) {
            inLatent <- c(latent1[sample1 == x], latent2[sample2
               ==
                x])[!duplicated(c(id1[sample1 == x], id2[sample2
```

```
                ==
            x]))]
        if (length(inLatent) == 0) return(0)
        return(sum(sample1 == x) * (mean(inLatent)^2 + mean(!
            inLatent)^2))
    }))
out$expectedOut <- sum(with(allNeighbors[outSelector, ],
    sapply(samples, function(x) {
        inLatent <- c(latent1[sample1 == x], latent2[sample2
            ==
            x])[!duplicated(c(id1[sample1 == x], id2[sample2
                ==
            x]))]
        outLatent <- c(latent1[sample1 != x], latent2[sample2
            !=
            x])[!duplicated(c(id1[sample1 != x], id2[sample2
                !=
            x]))]
        if (length(inLatent) == 0) return(0)
        return(sum(sample1 == x) * (mean(inLatent) * mean(
            outLatent) +
            mean(!inLatent) * mean(!outLatent)))
    }))
out$expectedAll <- sum(with(allNeighbors[allSelector, ],
    {
        allLatent <- c(latent1, latent2)[!duplicated(c(id1,
            id2))]
```

```
            return(length(latent1) * (mean(allLatent)^2 + mean(!
               allLatent)^2))
        }))
    return(out)
}, allNeighbors))

rownames(pairings) <- pairings$dist
```

To look for more matches than expected by random pairing between neighboring proviruses, we used a one sample Z-test of proportion to compare the observed number of matching pairs with the expected proportion of pairs.

```
combinations <- c(All = "All", `Between sample` = "Out", `Within
   sample` = "In")

lapply(combinations, function(x, pairing) {
    vars <- sprintf(c("observed%s", "expected%s", "total%s"),
        x)
    expectedProb <- pairing[, vars[2]]/pairing[, vars[3]]
    prop.test(pairing[, vars[1]], pairing[, vars[3]], p =
       expectedProb)
}, pairings["100", ])

## $All
##
##      1-sample proportions test with continuity correction
##
## data:  pairing[, vars[1]] out of pairing[, vars[3]], null
   probability expectedProb
```

```
## X-squared = 13.002, df = 1, p-value = 0.0003111
## alternative hypothesis: true p is not equal to 0.5000141
## 95 percent confidence interval:
##  0.5586837 0.6962353
## sample estimates:
##    p
## 0.63
##
##
## $`Between sample`
##
##      1-sample proportions test with continuity correction
##
## data:  pairing[, vars[1]] out of pairing[, vars[3]], null
   probability expectedProb
## X-squared = 0.21919, df = 1, p-value = 0.6397
## alternative hypothesis: true p is not equal to 0.4836763
## 95 percent confidence interval:
##  0.3570532 0.5572662
## sample estimates:
##         p
## 0.4554455
##
##
## $`Within sample`
##
##      1-sample proportions test with continuity correction
```

```
##
## data:  pairing[, vars[1]] out of pairing[, vars[3]], null
   probability expectedProb
## X-squared = 24.446, df = 1, p-value = 7.644e-07
## alternative hypothesis: true p is not equal to 0.5561437
## 95 percent confidence interval:
##  0.7140170 0.8776751
## sample estimates:
##         p
## 0.8080808
```

A.1.10 Compiling this document

This document was generated using R's Sweave function (`http://en.wikipedia.org/wiki/Sweave`). If you would like to regenerate this document, download Additional Files 2, 3 and 4 from Sherrill-Mix et al. [189] and make sure the files are all in the same directory and named AdditionalFile2.csv.gz, AdditionalFile3.csv.gz and AdditionalFile4.Rnw. Then compile by going to that directory and using the commands:
R CMD Sweave AdditionalFile4.Rnw
pdflatex AdditionalFile4.tex

Note that you will need R and LaTeX (and the R package glmnet if you would like to rerun the lasso regressions) installed.

BIBLIOGRAPHY

[1] Michael S Gottlieb, Howard M Schanker, Peng Thim Fan, Andrew Saxon, Joel D Weisman, and Irving Pozalski. Pneumocystis pneumonia—Los Angeles. *MMWR Morb Mortal Wkly Rep*, 30(21):250–252, Jun 1981. URL http://www.cdc.gov/mmwr/preview/mmwrhtml/june_5.htm.

[2] AE Friedman-Kien, L Laubenstein, M Marmor, K Hymes, J Green, A Ragaz, J Gottleib, F Muggia, R Demopoulos, and M Weintraub. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men–New York City and California. *MMWR Morb Mortal Wkly Rep*, 30(25):305–308, Jul 1981. URL http://www.ncbi.nlm.nih.gov/pubmed/6789108.

[3] K. B. Hymes, T. Cheung, J. B. Greene, N. S. Prose, A. Marcus, H. Ballard, D. C. William, and L. J. Laubenstein. Kaposi's sarcoma in homosexual men—a report of eight cases. *Lancet*, 2(8247):598–600, Sep 1981. doi: 10.1016/S0140-6736(81)92740-9. URL http://dx.doi.org/10.1016/S0140-6736(81)92740-9.

[4] H. Masur, M. A. Michelis, J. B. Greene, I. Onorato, R. A. Stouwe, R. S. Holzman, G. Wormser, L. Brettman, M. Lange, H. W. Murray, and S. Cunningham-Rundles. An outbreak of community-acquired Pneumocystis carinii pneumonia: initial manifestation of cellular immune dysfunction. *N Engl J Med*, 305(24):1431–1438, Dec 1981. doi: 10.1056/NEJM198112103052402. URL http://dx.doi.org/10.1056/NEJM198112103052402.

[5] F. P. Siegal, C. Lopez, G. S. Hammer, A. E. Brown, S. J. Kornfeld, J. Gold, J. Hassett, S. Z. Hirschman, C. Cunningham-Rundles, and B. R. Adelsberg. Severe acquired immunodeficiency in male homosexuals, manifested by chronic perianal ulcerative herpes simplex lesion. *N Engl J Med*, 305(24):1439–1444, Dec 1981. doi: 10.1056/NEJM198112103052403. URL http://dx.doi.org/10.1056/NEJM198112103052403.

[6] M. S. Gottlieb, R. Schroff, H. M. Schanker, J. D. Weisman, P. T. Fan, R. A. Wolf, and A. Saxon. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *N Engl J Med*, 305(24):1425–1431, Dec 1981. doi: 10.1056/NEJM198112103052401. URL http://dx.doi.org/10.1056/NEJM198112103052401.

[7] Y. Laor and R. A. Schwartz. Epidemiologic aspects of American Kaposi's sarcoma. *J Surg Oncol*, 12(4):299–303, Dec 1979. doi: 10.1002/jso.2930120403.

[8] M. B. Klein, F. A. Pereira, and I. Kantor. Kaposi Sarcoma complicating systemic lupus erythematosus treated with immunosuppression. *Arch Dermatol*, 110(4):602–604, Oct 1974. doi: 10.1001/archderm.1974.01630100058014. URL http://dx.doi.org/10.1001/archderm.1974.01630100058014.

[9] B. D. Myers, E. Kessler, J. Levi, A. Pick, J. B. Rosenfeld, and P. Tikvah. Kaposi sarcoma in kidney transplant recipients. *Arch Intern Med*, 133(2):307–311, Feb 1974. doi: 10.1001/archinte.1974.00320140145017. URL htttp://dx.doi.org/10.1001/archinte.1974.00320140145017.

[10] S. B. Kapadia and J. R. Krause. Kaposi's sarcoma after long-term alkylating agent therapy for multiple myeloma. *South Med J*, 70(8):1011–1013, Aug 1977. URL http://www.ncbi.nlm.nih.gov/pubmed/887963.

[11] B. Safai and R. A. Good. Kaposi's sarcoma: a review and recent developments. *CA Cancer J Clin*, 31(1):2–12, 1981. doi: 10.3322/canjclin.31.1.2. URL http://dx.doi.org/10.3322/canjclin.31.1.2.

[12] Y. Chang, E. Cesarman, M. S. Pessin, F. Lee, J. Culpepper, D. M. Knowles, and P. S. Moore. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science*, 266(5192):1865–1869, Dec 1994. doi: 10.1126/science.7997879. URL http://dx.doi.org/10.1126/science.7997879.

[13] F. Sitas, H. Carrara, V. Beral, R. Newton, G. Reeves, D. Bull, U. Jentsch, R. Pacella-Norman, D. Bourboulia, D. Whitby, C. Boshoff, and R. Weiss. Antibodies against human herpesvirus 8 in black South African patients with cancer. *N Engl J Med*, 340(24):1863–1871, Jun 1999. doi: 10.1056/NEJM199906173402403. URL http://dx.doi.org/10.1056/NEJM199906173402403.

[14] B. A. Burke and R. A. Good. Pneumocystis carinii infection. *Medicine (Baltimore)*, 52(1):23–51, Jan 1973. URL http://journals.lww.com/md-journal/Citation/1973/01000/PNEUMOCYSTIS_CARINII_INFECTION_.2.aspx.

[15] W. T. Hughes. Pneumocystis carinii pneumonia. *N Engl J Med*, 297(25):1381–1383, Dec 1977. doi: 10.1056/NEJM197712222972505. URL http://dx.doi.org/10.1056/NEJM197712222972505.

[16] James R. Stringer, Charles B. Beard, and Robert F. Miller. Spelling Pneumocystis jirovecii. *Emerg Infect Dis*, 15(3):506, Mar 2009. doi: 10.3201/eid1503.081060. URL http://dx.doi.org/10.3201/eid1503.081060.

[17] J. Gerstoft, A. Malchow-Møller, I. Bygbjerg, E. Dickmeiss, C. Enk, P. Halberg, S. Haahr, M. Jacobsen, K. Jensen, J. Mejer, J. O. Nielsen, H. K. Thomsen, J. Søndergaard, and I. Lorenzen. Severe acquired immunodeficiency in European homosexual men. *Br Med J (Clin Res Ed)*, 285(6334):17–19, Jul 1982. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1499130/.

[18] H. Masur, M. A. Michelis, G. P. Wormser, S. Lewin, J. Gold, M. L. Tapper, J. Giron, C. W. Lerner, D. Armstrong, U. Setia, J. A. Sender, R. S. Siebken, P. Nicholas, Z. Arlen, S. Maayan, J. A. Ernst, F. P. Siegal, and S. Cunningham-Rundles. Opportunistic infection in previously healthy women. Initial manifestations of a

community-acquired cellular immunodeficiency. *Ann Intern Med*, 97(4):533–539, Oct 1982. URL http://www.ncbi.nlm.nih.gov/pubmed/6982014.

[19] Arthur Ammann, M Cowan, D Wara, H Goldman, H Perkins, R Lanzerotti, J Gullett, A Duff, S Dritz, and J Chin. Possible transfusion-associated acquired immune deficiency syndrome (AIDS) — California. *MMWR Morb Mortal Wkly Rep*, 31(48):652–654, Dec 1982. URL http://www.cdc.gov/mmwr/preview/mmwrhtml/00001203.htm.

[20] NJ Ehrenkranz, J Rubini, R Gunn, CR Horsburgh, T Collins, U Hasiba, W Hathaway, W Doig, R Hopkins, and J Elliott. Pneumocystis carinii pneumonia among persons with hemophilia A. *MMWR Morb Mortal Wkly Rep*, 31(27):365–367, Jul 1982. URL http://www.cdc.gov/mmwr/preview/mmwrhtml/00001126.htm.

[21] J. B. Greene, G. S. Sidhu, S. Lewin, J. F. Levine, H. Masur, M. S. Simberkoff, P. Nicholas, R. C. Good, S. B. Zolla-Pazner, A. A. Pollock, M. L. Tapper, and R. S. Holzman. Mycobacterium avium-intracellulare: a cause of disseminated life-threatening infection in homosexuals and drug abusers. *Ann Intern Med*, 97(4): 539–546, Oct 1982. URL http://annals.org/article.aspx?articleid=695936.

[22] S Fannin, MS Gottlieb, JD Weisman, E Rogolsky, T Prendergast, J Chin, AE Friedman-Kien, L Laubenstein, S Friedman, and R Rothenberg. A cluster of Kaposi's sarcoma and Pneumocystis carinii pneumonia among homosexual male residents of Los Angeles and Orange Counties, California. *MMWR Morb Mortal Wkly Rep*, 31(23):305–307, Jun 1982. URL http://www.cdc.gov/mmwr/preview/mmwrhtml/00001114.htm.

[23] C Harris, C Butkus Small, G Friedland, R Klein, B Moll, E Emeson, I Spigland, N Steigbigel, R Reiss, S Friedman, and R Rothenberg. Immunodeficiency among female sexual partners of males with acquired immune deficiency syndrome (AIDS) — New York. *MMWR Morb Mortal Wkly Rep*, 31(52):697–698, Jan 1983. URL http://www.cdc.gov/mmwr/preview/mmwrhtml/00001221.htm.

[24] F. Barré-Sinoussi, J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vézinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):868–871, May 1983. URL http://www.sciencemag.org/cgi/pmidlookup?view=long&pmid=6189183.

[25] M. Popovic, M. G. Sarngadharan, E. Read, and R. C. Gallo. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science*, 224(4648):497–500, May 1984. doi: 10.1126/science.6200935. URL http://dx.doi.org/10.1126/science.6200935.

[26] R. C. Gallo, S. Z. Salahuddin, M. Popovic, G. M. Shearer, M. Kaplan, B. F. Haynes, T. J. Palker, R. Redfield, J. Oleske, and B. Safai. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*, 224(4648):500–503, May 1984. doi: 10.1126/science.6200936. URL http://dx.doi.org/10.1126/science.6200936.

[27] M. G. Sarngadharan, M. Popovic, L. Bruch, J. Schüpbach, and R. C. Gallo. Antibodies reactive with human T-lymphotropic retroviruses (HTLV-III) in the serum of patients with AIDS. *Science*, 224(4648):506–508, May 1984. doi: 10.1126/science.6324345.

[28] B. Safai, M. G. Sarngadharan, J. E. Groopman, K. Arnett, M. Popovic, A. Sliski, J. Schüpbach, and R. C. Gallo. Seroepidemiological studies of human T-lymphotropic retrovirus type iii in acquired immunodeficiency syndrome. *Lancet*, 1(8392):1438–1440, Jun 1984. doi: 10.1016/S0140-6736(84)91933-0. URL http://dx.doi.org/10.1016/S0140-6736(84)91933-0.

[29] S. Wain-Hobson, P. Sonigo, O. Danos, S. Cole, and M. Alizon. Nucleotide sequence of the AIDS virus, LAV. *Cell*, 40(1):9–17, Jan 1985. doi: 10.1016/0092-8674(85)90303-4. URL http://dx.doi.org/10.1016/0092-8674(85)90303-4.

[30] F. Gao, E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature*, 397(6718):436–441, Feb 1999. doi: 10.1038/17130. URL http://dx.doi.org/10.1038/17130.

[31] C. Gélinas and H. M. Temin. Nondefective spleen necrosis virus-derived vectors define the upper size limit for packaging reticuloendotheliosis viruses. *Proc Natl Acad Sci USA*, 83(23):9211–9215, Dec 1986. URL http://www.pnas.org/content/83/23/9211.abstract.

[32] S. A. Herman and J. M. Coffin. Efficient packaging of readthrough RNA in ALV: implications for oncogene transduction. *Science*, 236(4803):845–848, May 1987. doi: 10.1126/science.3033828. URL http://dx.doi.org/10.1126/science.3033828.

[33] N. H. Shin, D. Hartigan-O'Connor, J. K. Pfeiffer, and A. Telesnitsky. Replication of lengthened Moloney murine leukemia virus genomes is impaired at multiple stages. *J Virol*, 74(6):2694–2702, Mar 2000. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC111759/.

[34] C. Martin Stoltzfus. Chapter 1. Regulation of HIV-1 alternative RNA splicing and its role in virus replication. *Adv Virus Res*, 74:1–40, 2009. doi: 10.1016/S0065-3527(09)74001-1. URL http://dx.doi.org/10.1016/S0065-3527(09)74001-1.

[35] S. Y. Kim, R. Byrn, J. Groopman, and D. Baltimore. Temporal aspects of DNA and RNA synthesis during human immunodeficiency virus infection: evidence for differential gene expression. *J Virol*, 63(9):3708–3713, Sep 1989. URL http://jvi.asm.org/cgi/content/abstract/63/9/3708.

[36] R. J. Pomerantz, D. Trono, M. B. Feinberg, and D. Baltimore. Cells nonproductively infected with HIV-1 exhibit an aberrant pattern of viral RNA expression: a molecular model for latency. *Cell*, 61(7):1271–1276, Jun 1990. doi: 10.1016/0092-8674(90)90691-7. URL http://dx.doi.org/10.1016/0092-8674(90)90691-7.

[37] Abraham L Brass, Derek M Dykxhoorn, Yair Benita, Nan Yan, Alan Engelman, Ramnik J Xavier, Judy Lieberman, and Stephen J Elledge. Identification of host proteins required for HIV infection through a functional genomic screen. *Science*, 319(5865):921–926, Feb 2008. doi: 10.1126/science.1152725. URL http://dx.doi.org/10.1126/science.1152725.

[38] Renate König, Yingyao Zhou, Daniel Elleder, Tracy L Diamond, Ghislain M C Bonamy, Jeffrey T Irelan, Chih-Yuan Chiang, Buu P Tu, Paul D De Jesus, Caroline E Lilley, Shannon Seidel, Amanda M Opaluch, Jeremy S Caldwell, Matthew D Weitzman, Kelli L Kuhen, Sourav Bandyopadhyay, Trey Ideker, Anthony P Orth, Loren J Miraglia, Frederic D Bushman, John A Young, and Sumit K Chanda. Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, 135(1):49–60, Oct 2008. doi: 10.1016/j.cell.2008.07.032. URL http://dx.doi.org/10.1016/j.cell.2008.07.032.

[39] Frederic D Bushman, Nirav Malani, Jason Fernandes, Ivn D'Orso, Gerard Cagney, Tracy L Diamond, Honglin Zhou, Daria J Hazuda, Amy S Espeseth, Renate Knig, Sourav Bandyopadhyay, Trey Ideker, Stephen P Goff, Nevan J Krogan, Alan D Frankel, John A T Young, and Sumit K Chanda. Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog*, 5(5):e1000437, May 2009. doi: 10.1371/journal.ppat.1000437. URL http://dx.doi.org/10.1371/journal.ppat.1000437.

[40] Stefanie Jäger, Peter Cimermancic, Natali Gulbahce, Jeffrey R Johnson, Kathryn E McGovern, Starlynn C Clarke, Michael Shales, Gaelle Mercenne, Lars Pache, Kathy Li, Hilda Hernandez, Gwendolyn M Jang, Shoshannah L Roth, Eyal Akiva, John Marlett, Melanie Stephens, Ivn D'Orso, Jason Fernandes, Marie Fahey, Cathal Mahon, Anthony J O'Donoghue, Aleksandar Todorovic, John H Morris, David A Maltby, Tom Alber, Gerard Cagney, Frederic D Bushman, John A Young, Sumit K Chanda, Wesley I Sundquist, Tanja Kortemme, Ryan D Hernandez, Charles S Craik, Alma Burlingame, Andrej Sali, Alan D Frankel, and Nevan J Krogan. Global landscape of HIV-human protein complexes. *Nature*, 481(7381):365–370, Jan 2012. doi: 10.1038/nature10719. URL http://dx.doi.org/10.1038/nature10719.

[41] Anju Bansal, Jonathan Carlson, Jiyu Yan, Olusimidele T Akinsiku, Malinda Schaefer, Steffanie Sabbaj, Anne Bet, David N Levy, Sonya Heath, Jianming Tang, Richard A Kaslow, Bruce D Walker, Thumbi Ndung'u, Philip J Goulder, David Heckerman, Eric Hunter, and Paul A Goepfert. CD8 T cell response and evolutionary pressure to HIV-1 cryptic epitopes derived from antisense transcription. *J Exp Med*, 207(1):51–59, Jan 2010. doi: 10.1084/jem.20092060. URL http://dx.doi.org/10.1084/jem.20092060.

[42] Takeshi Fukuhara, Takamitsu Hosoya, Saki Shimizu, Kengo Sumi, Takako Oshiro, Yoshiyuki Yoshinaka, Masaaki Suzuki, Naoki Yamamoto, Leonore A Herzenberg, Leonard A Herzenberg, and Masatoshi Hagiwara. Utilization of host SR protein kinases and RNA-splicing machinery during viral replication. *Proc Natl Acad Sci USA*, 103(30):11329–11333, Jul 2006. doi: 10.1073/pnas.0604616103. URL http://dx.doi.org/10.1073/pnas.0604616103.

[43] Nadia Bakkour, Yea-Lih Lin, Sophie Maire, Lilia Ayadi, Florence Mahuteau-Betzer, Chi Hung Nguyen, Clment Mettling, Pierre Portales, David Grierson, Benoit Chabot, Philippe Jeanteur, Christiane Branlant, Pierre Corbeau, and Jamal Tazi. Small-molecule inhibition of HIV pre-mRNA splicing as a novel antiretroviral therapy to overcome drug resistance. *PLoS Pathog*, 3(10):1530–1539, Oct 2007. doi: 10.1371/journal.ppat.0030159. URL http://dx.doi.org/10.1371/journal.ppat.0030159.

[44] Maria B Asparuhova, Gabriela Marti, Songkai Liu, Fatima Serhan, Didier Trono, and Daniel Schmperli. Inhibition of HIV-1 multiplication by a modified U7 snRNA inducing Tat and Rev exon skipping. *J Gene Med*, 9(5):323–334, May 2007. doi: 10.1002/jgm.1027. URL http://dx.doi.org/10.1002/jgm.1027.

[45] Dibyakanti Mandal, Zehua Feng, and C. Martin Stoltzfus. Excessive RNA splicing and inhibition of HIV-1 replication induced by modified U1 small nuclear RNAs. *J Virol*, 84(24):12790–12800, Dec 2010. doi: 10.1128/JVI.01257-10. URL http://dx.doi.org/10.1128/JVI.01257-10.

[46] T. O. Tange, T. H. Jensen, and J. Kjems. In vitro interaction between human immunodeficiency virus type 1 Rev protein and splicing factor ASF/SF2-associated protein, p32. *J Biol Chem*, 271(17):10066–10072, Apr 1996. doi: 10.1074/jbc.271.17.10066. URL http://dx.doi.org/10.1074/jbc.271.17.10066.

[47] Reem Berro, Kylene Kehn, Cynthia de la Fuente, Anne Pumfery, Richard Adair, John Wade, Anamaris M Colberg-Poley, John Hiscott, and Fatah Kashanchi. Acetylated Tat regulates human immunodeficiency virus type 1 splicing through its interaction with the splicing regulator p32. *J Virol*, 80(7):3189–3204, Apr 2006. doi: 10.1128/JVI.80.7.3189-3204.2006. URL http://dx.doi.org/10.1128/JVI.80.7.3189-3204.2006.

[48] Jens Bohne, Axel Schambach, and Daniela Zychlinski. New way of regulating alternative splicing in retroviruses: the promoter makes a difference. *J Virol*,

81(7):3652–3656, Apr 2007. doi: 10.1128/JVI.02105-06. URL http://dx.doi.org/10.1128/JVI.02105-06.

[49] Joseph A Jablonski, Antonio L Amelio, Mauro Giacca, and Massimo Caputi. The transcriptional transactivator Tat selectively regulates viral splicing. *Nucleic Acids Res*, 38(4):1249–1260, Mar 2010. doi: 10.1093/nar/gkp1105. URL http://dx.doi.org/10.1093/nar/gkp1105.

[50] Madoka Kuramitsu, Chieko Hashizume, Norio Yamamoto, Akihiko Azuma, Masakazu Kamata, Naoki Yamamoto, Yoshimasa Tanaka, and Yoko Aida. A novel role for Vpr of human immunodeficiency virus type 1 as a regulator of the splicing of cellular pre-mRNA. *Microbes Infect*, 7(9-10):1150–1160, Jul 2005. doi: 10.1016/j.micinf.2005.03.022. URL http://dx.doi.org/10.1016/j.micinf.2005.03.022.

[51] Chieko Hashizume, Madoka Kuramitsu, Xianfeng Zhang, Terue Kurosawa, Masakazu Kamata, and Yoko Aida. Human immunodeficiency virus type 1 Vpr interacts with spliceosomal protein SAP145 to mediate cellular pre-mRNA splicing inhibition. *Microbes Infect*, 9(4):490–497, Apr 2007. doi: 10.1016/j.micinf.2007.01.013. URL http://dx.doi.org/10.1016/j.micinf.2007.01.013.

[52] Naama M Kopelman, Doron Lancet, and Itai Yanai. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat Genet*, 37(6):588–589, Jun 2005. doi: 10.1038/ng1575. URL http://dx.doi.org/10.1038/ng1575.

[53] Yi Xing and Christopher Lee. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc Natl Acad Sci USA*, 102(38):13526–13531, Sep 2005. doi: 10.1073/pnas.0501213102. URL http://dx.doi.org/10.1073/pnas.0501213102.

[54] Zhixi Su, Jianmin Wang, Jun Yu, Xiaoqiu Huang, and Xun Gu. Evolution of alternative splicing after gene duplication. *Genome Res*, 16(2):182–189, Feb 2006. doi: 10.1101/gr.4197006. URL http://dx.doi.org/10.1101/gr.4197006.

[55] Fiona L Watson, Roland Pttmann-Holgado, Franziska Thomas, David L Lamar, Michael Hughes, Masahiro Kondo, Vivienne I Rebel, and Dietmar Schmucker. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science*, 309(5742):1874–1878, Sep 2005. doi: 10.1126/science.1116887. URL http://dx.doi.org/10.1126/science.1116887.

[56] V. W. Pollard and M. H. Malim. The HIV-1 Rev protein. *Annu Rev Microbiol*, 52:491–532, 1998. doi: 10.1146/annurev.micro.52.1.491. URL http://dx.doi.org/10.1146/annurev.micro.52.1.491.

[57] D. F. Purcell and M. A. Martin. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expres-

sion, replication, and infectivity. *J Virol*, 67(11):6365–6378, Nov 1993. URL http://jvi.asm.org/cgi/content/abstract/67/11/6365?ijkey=49e1b7e6cfa47c60983887d47e00b14615372ec3&keytype2=tf_ipsecsha.

[58] M. E. Klotman, S. Kim, A. Buchbinder, A. DeRossi, D. Baltimore, and F. Wong-Staal. Kinetics of expression of multiply spliced RNA in early human immunodeficiency virus type 1 infection of lymphocytes and monocytes. *Proc Natl Acad Sci USA*, 88(11):5011–5015, Jun 1991. URL http://www.pnas.org/content/88/11/5011.abstract.

[59] D. M. Benko, S. Schwartz, G. N. Pavlakis, and B. K. Felber. A novel human immunodeficiency virus type 1 protein, tev, shares sequences with tat, env, and rev proteins. *J Virol*, 64(6):2505–2518, Jun 1990. URL http://jvi.asm.org/cgi/content/abstract/64/6/2505?ijkey=275e102c2bc0416e1d587fae36888f656248c0fd&keytype2=tf_ipsecsha.

[60] K. Fujita, J. Silver, and K. Peden. Changes in both gp120 and gp41 can account for increased growth potential and expanded host range of human immunodeficiency virus type 1. *J Virol*, 66(7):4445–4451, Jul 1992. URL http://jvi.asm.org/content/66/7/4445.short.

[61] R. M. McAllister, M. B. Gardner, A. E. Greene, C. Bradt, W. W. Nichols, and B. H. Landing. Cultivation in vitro of cells derived from a human osteosarcoma. *Cancer*, 27(2):397–402, Feb 1971. URL http://onlinelibrary.wiley.com/doi/10.1002/1097-0142%28197102%2927:2%3C397::AID-CNCR2820270224%3E3.0.CO;2-X/abstract.

[62] Tony Kwan, David Benovoy, Christel Dias, Scott Gurd, David Serre, Harry Zuzan, Tyson A Clark, Anthony Schweitzer, Michelle K Staples, Hui Wang, John E Blume, Thomas J Hudson, Rob Sladek, and Jacek Majewski. Heritability of alternative splicing in the human genome. *Genome Res*, 17(8):1210–1218, Aug 2007. doi: 10.1101/gr.6281007. URL http://dx.doi.org/10.1101/gr.6281007.

[63] Jeremy Hull, Susana Campino, Kate Rowlands, Man-Suen Chan, Richard R Copley, Martin S Taylor, Kirk Rockett, Gareth Elvidge, Brendan Keating, Julian Knight, and Dominic Kwiatkowski. Identification of common genetic variation that modulates alternative splicing. *PLoS Genet*, 3(6):e99, Jun 2007. doi: 10.1371/journal.pgen.0030099. URL http://dx.doi.org/10.1371/journal.pgen.0030099.

[64] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456:470–476, 2008. doi: 10.1038/nature07509. URL http://www.nature.com/nature/journal/v456/n7221/abs/nature07509.html.

[65] Yoseph Barash, John A Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J Blencowe, and Brendan J Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010. doi: 10.1038/nature09000. URL http://dx.doi.org/10.1038/nature09000.

[66] Maryanne T Vahey, Martin E Nau, Linda L Jagodzinski, Jake Yalley-Ogunro, Michele Taubman, Nelson L Michael, and Mark G Lewis. Impact of viral infection on the gene expression profiles of proliferating normal human peripheral blood mononuclear cells infected with HIV type 1 RF. *AIDS Res Hum Retroviruses*, 18 (3):179–192, Feb 2002. doi: 10.1089/08892220252781239. URL http://dx.doi.org/10.1089/08892220252781239.

[67] Anglique B van 't Wout, Ginger K Lehrman, Svetlana A Mikheeva, Gemma C O'Keeffe, Michael G Katze, Roger E Bumgarner, Gary K Geiss, and James I Mullins. Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4(+)-T-cell lines. *J Virol*, 77(2):1392–1402, Jan 2003. doi: 10.1128/JVI.77.2.1392-1402.2003. URL http://dx.doi.org/10.1128/JVI.77.2.1392-1402.2003.

[68] Richard Mitchell, Chih-Yuan Chiang, Charles Berry, and Frederic Bushman. Global analysis of cellular transcription following infection with an HIV-based vector. *Mol Ther*, 8(4):674–687, Oct 2003. doi: 10.1016/S1525-0016(03)00215-6. URL http://dx.doi.org/10.1016/S1525-0016(03)00215-6.

[69] Margalida Rotger, Kristen K Dang, Jacques Fellay, Erin L Heinzen, Sheng Feng, Patrick Descombes, Kevin V Shianna, Dongliang Ge, Huldrych F Gnthard, David B Goldstein, Amalio Telenti, Swiss HIV Cohort Study, and Center for HIV/AIDS Vaccine Immunology. Genome-wide mRNA expression correlates of viral control in CD4+ T-cells from HIV-1-infected individuals. *PLoS Pathog*, 6(2):e1000781, Feb 2010. doi: 10.1371/journal.ppat.1000781. URL http://dx.doi.org/10.1371/journal.ppat.1000781.

[70] Stewart T Chang, Pavel Sova, Xinxia Peng, Jeffrey Weiss, G. Lynn Law, Robert E Palermo, and Michael G Katze. Next-generation sequencing reveals HIV-1-mediated suppression of T cell activation and RNA processing and regulation of noncoding RNA expression in a CD4+ T cell line. *MBio*, 2(5), 2011. doi: 10.1128/mBio.00134-11. URL http://dx.doi.org/10.1128/mBio.00134-11.

[71] Ryan Tewhey, Jason B Warner, Masakazu Nakano, Brian Libby, Martina Medkova, Patricia H David, Steve K Kotsopoulos, Michael L Samuels, J. Brian Hutchison, Jonathan W Larson, Eric J Topol, Michael P Weiner, Olivier Harismendy, Jeff Olson, Darren R Link, and Kelly A Frazer. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol*, 27(11):1025–1031, Nov 2009. doi: 10.1038/nbt.1583. URL http://dx.doi.org/10.1038/nbt.1583.

[72] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: an assessment of technical reproducibility and comparison

with gene expression arrays. *Genome Res*, 18(9):1509–1517, Sep 2008. doi: 10.1101/gr.079558.108. URL http://dx.doi.org/10.1101/gr.079558.108.

[73] Ryan Morin, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor Pugh, Helen McDonald, Richard Varhol, Steven Jones, and Marco Marra. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques*, 45(1):81–94, Jul 2008. doi: 10.2144/000112900. URL http://dx.doi.org/10.2144/000112900.

[74] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323:133–138, 2009. doi: 10.1126/science.1162986. URL http://www.sciencemag.org/cgi/content/abstract/323/5910/133.

[75] Malinda Schaefer, M Brown, W Kilembe, S Allen, Y Guo, E Hunter, and E Paxinos. Single-molecule complete HIV-1 genome sequencing from 2 linked transmission pairs. In *Conference on Retroviruses and Opportunistic Infections*, 2012.

[76] Emanuele Buratti and Francisco E Baralle. Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol Cell Biol*, 24(24):10505–10514, Dec 2004. doi: 10.1128/MCB.24.24.10505-10514.2004. URL http://dx.doi.org/10.1128/MCB.24.24.10505-10514.2004.

[77] Joseph A Jablonski, Emanuele Buratti, Cristiana Stuani, and Massimo Caputi. The secondary structure of the human immunodeficiency virus type 1 transcript modulates viral splicing and infectivity. *J Virol*, 82(16):8038–8050, Aug 2008. doi: 10.1128/JVI.00721-08. URL http://dx.doi.org/10.1128/JVI.00721-08.

[78] Peter J Shepard and Klemens J Hertel. Conserved RNA secondary structures promote alternative splicing. *RNA*, 14(8):1463–1469, Aug 2008. doi: 10.1261/rna.1069408. URL http://dx.doi.org/10.1261/rna.1069408.

[79] Mariano Alló, Valeria Buggiano, Juan P Fededa, Ezequiel Petrillo, Ignacio Schor, Manuel de la Mata, Eneritz Agirre, Mireya Plass, Eduardo Eyras, Sherif Abou Elela, Roscoe Klinck, Benoit Chabot, and Alberto R Kornblihtt. Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nat Struct Mol Biol*, 16(7):717–724, Jul 2009. doi: 10.1038/nsmb.1620. URL http://dx.doi.org/10.1038/nsmb.1620.

[80] Hagen Tilgner, Christoforos Nikolaou, Sonja Althammer, Michael Sammeth, Miguel Beato, Juan Valcrcel, and Roderic Guig. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol*, 16(9):996–1001, Sep 2009. doi: 10.1038/nsmb.1658. URL http://dx.doi.org/10.1038/nsmb.1658.

[81] Schraga Schwartz, Eran Meshorer, and Gil Ast. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, 16(9):990–995, Sep 2009. doi: 10.1038/nsmb.1659. URL http://dx.doi.org/10.1038/nsmb.1659.

[82] Tara L Crabb, Bianca J Lam, and Klemens J Hertel. Retention of spliceosomal components along ligated exons ensures efficient removal of multiple introns. *RNA*, 16(9):1786–1796, Sep 2010. doi: 10.1261/rna.2186510. URL http://dx.doi.org/10.1261/rna.2186510.

[83] Kazuhiko Takahara, Ulrike Schwarze, Yasutada Imamura, Guy G Hoffman, Helga Toriello, Lynne T Smith, Peter H Byers, and Daniel S Greenspan. Order of intron removal influences multiple splice outcomes, including a two-exon skip, in a COL5A1 acceptor-site mutation that results in abnormal pro-alpha1(V) N-propeptides and Ehlers-Danlos syndrome type I. *Am J Hum Genet*, 71(3):451–465, Sep 2002. doi: 10.1086/342099. URL http://dx.doi.org/10.1086/342099.

[84] Manuel de la Mata, Celina Lafaille, and Alberto R Kornblihtt. First come, first served revisited: factors affecting the same alternative splicing event have different effects on the relative rates of intron removal. *RNA*, 16(5):904–912, May 2010. doi: 10.1261/rna.1993510. URL http://dx.doi.org/10.1261/rna.1993510.

[85] A. M. Zahler, K. M. Neugebauer, W. S. Lane, and M. B. Roth. Distinct functions of SR proteins in alternative pre-mRNA splicing. *Science*, 260(5105):219–222, Apr 1993. doi: 10.1126/science.8385799. URL http://dx.doi.org/10.1126/science.8385799.

[86] C. W. Smith and J. Valcárcel. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci*, 25(8):381–388, Aug 2000. doi: 10.1016/S0968-0004(00)01604-2. URL http://dx.doi.org/10.1016/S0968-0004(00)01604-2.

[87] Jernej Ule, Giovanni Stefani, Aldo Mele, Matteo Ruggiu, Xuning Wang, Bahar Taneri, Terry Gaasterland, Benjamin J Blencowe, and Robert B Darnell. An RNA map predicting Nova-dependent splicing regulation. *Nature*, 444(7119):580–586, Nov 2006. doi: 10.1038/nature05304. URL http://dx.doi.org/10.1038/nature05304.

[88] Hui Yuan Xiong, Yoseph Barash, and Brendan J Frey. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics*, 27(18):2554–2562, Sep 2011. doi: 10.1093/bioinformatics/btr444. URL http://dx.doi.org/10.1093/bioinformatics/btr444.

[89] Joshua T Witten and Jernej Ule. Understanding splicing regulation through RNA splicing maps. *Trends Genet*, 27(3):89–97, Mar 2011. doi: 10.1016/j.tig.2010.12.001. URL http://dx.doi.org/10.1016/j.tig.2010.12.001.

[90] M. M. O'Reilly, M. T. McNally, and K. L. Beemon. Two strong 5' splice sites and competing, suboptimal 3' splice sites involved in alternative splicing of human immunodeficiency virus type 1 RNA. *Virology*, 213(2):373–385, Nov 1995. doi: 10.1006/viro.1995.0010. URL http://dx.doi.org/10.1006/viro.1995.0010.

[91] B. A. Amendt, D. Hesslein, L. J. Chang, and C. M. Stoltzfus. Presence of negative and positive cis-acting RNA splicing elements within and flanking the first tat coding exon of human immunodeficiency virus type 1. *Mol Cell Biol*, 14(6):3960–3970, Jun 1994. doi: 10.1128/MCB.14.6.3960. URL http://dx.doi.org/10.1128/MCB.14.6.3960.

[92] Jeffrey D Levengood, Carrie Rollins, Clay H J Mishler, Charles A Johnson, Grace Miner, Prashant Rajan, Brent M Znosko, and Blanton S Tolbert. Solution structure of the HIV-1 exon splicing silencer 3. *J Mol Biol*, 415(4):680–698, Jan 2012. doi: 10.1016/j.jmb.2011.11.034. URL http://dx.doi.org/10.1016/j.jmb.2011.11.034.

[93] Massimo Caputi, Marcel Freund, Susanne Kammler, Corinna Asang, and Heiner Schaal. A bidirectional SF2/ASF- and SRp40-dependent splicing enhancer regulates human immunodeficiency virus type 1 rev, env, vpu, and nef gene expression. *J Virol*, 78(12):6517–6526, Jun 2004. doi: 10.1128/JVI.78.12.6517-6526.2004. URL http://dx.doi.org/10.1128/JVI.78.12.6517-6526.2004.

[94] Corinna Asang, Ilona Hauber, and Heiner Schaal. Insights into the selective activation of alternatively used splice acceptors by the human immunodeficiency virus type-1 bidirectional splicing enhancer. *Nucleic Acids Res*, 36(5):1450–1463, Mar 2008. doi: 10.1093/nar/gkm1147. URL http://dx.doi.org/10.1093/nar/gkm1147.

[95] T. O. Tange, C. K. Damgaard, S. Guth, J. Valcrcel, and J. Kjems. The hnRNP A1 protein regulates HIV-1 tat splicing via a novel intron silencer element. *EMBO J*, 20(20):5748–5758, Oct 2001. doi: 10.1093/emboj/20.20.5748. URL http://dx.doi.org/10.1093/emboj/20.20.5748.

[96] Anna Tranell, Eva Maria Feny, and Stefan Schwartz. Serine- and arginine-rich proteins 55 and 75 (SRp55 and SRp75) induce production of HIV-1 vpr mRNA by inhibiting the 5'-splice site of exon 3. *J Biol Chem*, 285(41):31537–31547, Oct 2010. doi: 10.1074/jbc.M109.077453. URL http://dx.doi.org/10.1074/jbc.M109.077453.

[97] C. Martin Stoltzfus and Joshua M Madsen. Role of viral splicing elements and cellular RNA binding proteins in regulation of HIV-1 alternative RNA splicing.

Curr HIV Res, 4(1):43–55, Jan 2006. doi: 10.2174/157016206775197655. URL http://dx.doi.org/10.2174/157016206775197655.

[98] P. Legrain and M. Rosbash. Some cis- and trans-acting mutants for splicing target pre-mRNA to the cytoplasm. *Cell*, 57(4):573–583, May 1989. doi: 10.1016/0092-8674(89)90127-X. URL http://dx.doi.org/10.1016/0092-8674(89)90127-X.

[99] U. Fischer, S. Meyer, M. Teufel, C. Heckel, R. Lhrmann, and G. Rautmann. Evidence that HIV-1 Rev directly promotes the nuclear export of unspliced RNA. *EMBO J*, 13(17):4105–4112, Sep 1994. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC395333/?tool=pmcentrez.

[100] J. Sodroski, C. Rosen, F. Wong-Staal, S. Z. Salahuddin, M. Popovic, S. Arya, R. C. Gallo, and W. A. Haseltine. Trans-acting transcriptional regulation of human T-cell leukemia virus type III long terminal repeat. *Science*, 227(4683):171–173, Jan 1985. doi: 10.1126/science.2981427. URL http://dx.doi.org/10.1126/science.2981427.

[101] K. A. Jones and B. M. Peterlin. Control of RNA initiation and elongation at the HIV-1 promoter. *Annu Rev Biochem*, 63:717–743, 1994. doi: 10.1146/annurev.bi.63.070194.003441. URL http://dx.doi.org/10.1146/annurev.bi.63.070194.003441.

[102] T. W. McCloskey, M. Ott, E. Tribble, S. A. Khan, S. Teichberg, M. O. Paul, S. Pahwa, E. Verdin, and N. Chirmule. Dual role of HIV Tat in regulation of apoptosis in T cells. *J Immunol*, 158(2):1014–1019, Jan 1997. URL http://www.jimmunol.org/content/158/2/1014.abstract.

[103] Grant R Campbell, Eddy Pasquier, Jennifer Watkins, Veronique Bourgarel-Rey, Vincent Peyrot, Didier Esquieu, Pascale Barbier, Jean de Mareuil, Diane Braguer, Pontiano Kaleebu, David L Yirrell, and Erwann P Loret. The glutamine-rich region of the HIV-1 Tat protein is involved in T-cell apoptosis. *J Biol Chem*, 279(46):48197–48204, Nov 2004. doi: 10.1074/jbc.M406195200. URL http://dx.doi.org/10.1074/jbc.M406195200.

[104] Heather B Miller, Timothy J Robinson, Raluca Gordn, Alexander J Hartemink, and Mariano A Garcia-Blanco. Identification of Tat-SF1 cellular targets by exon array analysis reveals dual roles in transcription and splicing. *RNA*, 17(4):665–674, Apr 2011. doi: 10.1261/rna.2462011. URL http://dx.doi.org/10.1261/rna.2462011.

[105] M. E. Rogel, L. I. Wu, and M. Emerman. The human immunodeficiency virus type 1 vpr gene prevents cell proliferation during chronic infection. *J Virol*, 69(2):882–888, Feb 1995. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC188655/.

[106] R. A. Fouchier, B. E. Meyer, J. H. Simon, U. Fischer, A. V. Albright, F. Gonzlez-Scarano, and M. H. Malim. Interaction of the human immunodeficiency virus type 1 Vpr protein with the nuclear pore complex. *J Virol*, 72(7):6004–6013, Jul 1998. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC110405/.

[107] Amelie K Gubitz, Wenqin Feng, and Gideon Dreyfuss. The SMN complex. *Exp Cell Res*, 296(1):51–56, May 2004. doi: 10.1016/j.yexcr.2004.03.022. URL http://dx.doi.org/10.1016/j.yexcr.2004.03.022.

[108] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, May 2010. doi: 10.1038/nbt.1621. URL http://dx.doi.org/10.1038/nbt.1621.

[109] Mark F Rogers, Julie Thomas, Anireddy Sn Reddy, and Asa Ben-Hur. SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol*, 13(1):R4, Jan 2012. doi: 10.1186/gb-2012-13-1-r4. URL http://dx.doi.org/10.1186/gb-2012-13-1-r4.

[110] A. Ryo, Y. Suzuki, K. Ichiyama, T. Wakatsuki, N. Kondoh, A. Hada, M. Yamamoto, and N. Yamamoto. Serial analysis of gene expression in HIV-1-infected T cell lines. *FEBS Lett*, 462(1-2):182–186, Nov 1999. doi: 10.1016/S0014-5793(99)01526-4. URL http://dx.doi.org/10.1016/S0014-5793(99)01526-4.

[111] Gregory Lefebvre, Sbastien Desfarges, Frdric Uyttebroeck, Miguel Muoz, Niko Beerenwinkel, Jacques Rougemont, Amalio Telenti, and Angela Ciuffi. Analysis of HIV-1 expression level and sense of transcription by high-throughput sequencing of the infected cell. *J Virol*, 85(13):6205–6211, Jul 2011. doi: 10.1128/JVI.00252-11. URL http://dx.doi.org/10.1128/JVI.00252-11.

[112] S. K. Arya, C. Guo, S. F. Josephs, and F. Wong-Staal. Trans-activator gene of human T-lymphotropic virus type III (HTLV-III). *Science*, 229(4708):69–73, Jul 1985. URL http://www.sciencemag.org/content/229/4708/69.long.

[113] S. Schwartz, B. K. Felber, D. M. Benko, E. M. Fenyö, and G. N. Pavlakis. Cloning and functional analysis of multiply spliced mRNA species of human immunodeficiency virus type 1. *J Virol*, 64(6):2519–2529, Jun 1990. URL http://jvi.asm.org/content/64/6/2519.abstract.

[114] Karen E Ocwieja, Scott Sherrill-Mix, Rithun Mukherjee, Rebecca Custers-Allen, Patricia David, Michael Brown, Susana Wang, Darren R Link, Jeff Olson, Kevin Travers, Eric Schadt, and Frederic D Bushman. Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing. *Nucleic Acids Res*, 40(20):10345–10355, Nov 2012. doi: 10.1093/nar/gks753. URL http://dx.doi.org/10.1093/nar/gks753.

[115] J. He, S. Choe, R. Walker, P. Di Marzio, D. O. Morgan, and N. R. Landau. Human immunodeficiency virus type 1 viral protein R (Vpr) arrests cells in the G2 phase of the cell cycle by inhibiting p34cdc2 activity. *J Virol*, 69(11):6705–6711, Nov 1995. URL http://jvi.asm.org/content/69/11/6705.short.

[116] J. B. Jowett, V. Planelles, B. Poon, N. P. Shah, M. L. Chen, and I. S. Chen. The human immunodeficiency virus type 1 vpr gene arrests infected T cells in the G2 + M phase of the cell cycle. *J Virol*, 69(10):6304–6313, Oct 1995. URL http://jvi.asm.org/content/69/11/6705.short.

[117] W. C. Goh, M. E. Rogel, C. M. Kinsey, S. F. Michael, P. N. Fultz, M. A. Nowak, B. H. Hahn, and M. Emerman. HIV-1 Vpr increases viral expression by manipulation of the cell cycle: a mechanism for selection of Vpr in vivo. *Nat Med*, 4(1):65–71, Jan 1998. doi: 10.1038/nm0198-065. URL http://dx.doi.org/10.1038/nm0198-065.

[118] R. A. Marciniak and P. A. Sharp. HIV-1 Tat protein promotes formation of more-processive elongation complexes. *EMBO J*, 10(13):4189–4196, Dec 1991. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC453171/.

[119] P. Wei, M. E. Garber, S. M. Fang, W. H. Fischer, and K. A. Jones. A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell*, 92(4):451–462, Feb 1998. doi: 10.1016/S0092-8674(00)80939-3. URL http://dx.doi.org/10.1016/S0092-8674(00)80939-3.

[120] S. Kanazawa, T. Okamoto, and B. M. Peterlin. Tat competes with CIITA for the binding to P-TEFb and blocks the expression of MHC class II genes in HIV infection. *Immunity*, 12(1):61–70, Jan 2000. doi: 10.1016/S1074-7613(00)80159-4. URL http://dx.doi.org/10.1016/S1074-7613(00)80159-4.

[121] Matjaz Barboric, Jasper H N. Yik, Nadine Czudnochowski, Zhiyuan Yang, Ruichuan Chen, Xavier Contreras, Matthias Geyer, B. Matija Peterlin, and Qiang Zhou. Tat competes with HEXIM1 to increase the active pool of P-TEFb for HIV-1 transcription. *Nucleic Acids Res*, 35(6):2003–2012, 2007. doi: 10.1093/nar/gkm063. URL http://dx.doi.org/10.1093/nar/gkm063.

[122] Siobhan K. O'Brien, Hong Cao, Robin Nathans, Akbar Ali, and Tariq M. Rana. P-TEFb kinase complex phosphorylates histone H1 to regulate expression of cellular and HIV-1 genes. *J Biol Chem*, 285(39):29713–29720, Sep 2010. doi: 10.1074/jbc.M110.125997. URL http://dx.doi.org/10.1074/jbc.M110.125997.

[123] Lisa Muniz, Sylvain Egloff, Bettina Ughy, Beáta E. Jády, and Tamás Kiss. Controlling cellular P-TEFb activity by the HIV-1 transcriptional transactivator Tat. *PLoS Pathog*, 6(10):e1001152, 2010. doi: 10.1371/journal.ppat.1001152. URL http://dx.doi.org/10.1371/journal.ppat.1001152.

[124] J. Corbeil, D. Sheeter, D. Genini, S. Rought, L. Leoni, P. Du, M. Ferguson, D. R. Masys, J. B. Welsh, J. L. Fink, R. Sasik, D. Huang, J. Drenkow, D. D. Richman, and T. Gingeras. Temporal gene regulation during HIV-1 infection of human CD4+ T cells. *Genome Res*, 11(7):1198–1204, Jul 2001. doi: 10.1101/gr.180201. URL http://dx.doi.org/10.1101/gr.180201.

[125] Christopher H. Woelk, Florence Ottones, Christine R. Plotkin, Pinyi Du, Christy D. Royer, Steffney E. Rought, Jean Lozach, Roman Sasik, Richard S. Kornbluth, Douglas D. Richman, and Jacques Corbeil. Interferon gene expression following HIV type 1 infection of monocyte-derived macrophages. *AIDS Res Hum Retroviruses*, 20(11):1210–1222, Nov 2004. doi: 10.1089/0889222042545009. URL http://dx.doi.org/10.1089/0889222042545009.

[126] Martin D. Hyrcza, Colin Kovacs, Mona Loutfy, Roberta Halpenny, Lawrence Heisler, Stuart Yang, Olivia Wilkins, Mario Ostrowski, and Sandy D. Der. Distinct transcriptional profiles in ex vivo CD4+ and CD8+ T cells are established early in human immunodeficiency virus type 1 infection and are characterized by a chronic interferon response as well as extensive transcriptional changes in CD8+ T cells. *J Virol*, 81(7):3477–3486, Apr 2007. doi: 10.1128/JVI.01552-06. URL http://dx.doi.org/10.1128/JVI.01552-06.

[127] Jing Qin Wu, Dominic E. Dwyer, Wayne B. Dyer, Yee Hwa Yang, Bin Wang, and Nitin K. Saksena. Transcriptional profiles in CD8+ T cells from HIV+ progressors on HAART are characterized by coordinated up-regulation of oxidative phosphorylation enzymes and interferon responses. *Virology*, 380(1):124–135, Oct 2008. doi: 10.1016/j.virol.2008.06.039. URL http://dx.doi.org/10.1016/j.virol.2008.06.039.

[128] Anthony J. Smith, Qingsheng Li, Stephen W. Wietgrefe, Timothy W. Schacker, Cavan S. Reilly, and Ashley T. Haase. Host genes associated with HIV-1 replication in lymphatic tissue. *J Immunol*, 185(9):5417–5424, Nov 2010. doi: 10.4049/jimmunol.1002197. URL http://dx.doi.org/10.4049/jimmunol.1002197.

[129] Michaël Imbeault, Katia Giguère, Michel Ouellet, and Michel J Tremblay. Exon level transcriptomic profiling of HIV-1-infected CD4(+) T cells reveals virus-induced genes and host environment favorable for viral replication. *PLoS Pathog*, 8(8):e1002861, Aug 2012. doi: 10.1371/journal.ppat.1002861. URL http://dx.doi.org/10.1371/journal.ppat.1002861.

[130] Pejman Mohammadi, Sbastien Desfarges, Istvn Bartha, Beda Joos, Nadine Zangger, Miguel Muoz, Huldrych F Gnthard, Niko Beerenwinkel, Amalio Telenti, and Angela Ciuffi. 24 hours in the life of HIV-1 in a T cell line. *PLoS Pathog*, 9(1):e1003161, Jan 2013. doi: 10.1371/journal.ppat.1003161. URL http://dx.doi.org/10.1371/journal.ppat.1003161.

[131] Xinxia Peng, Pavel Sova, Richard R. Green, Matthew J. Thomas, Marcus J.

Korth, Sean Proll, Jiabao Xu, Yanbing Cheng, Kang Yi, Li Chen, Zhiyu Peng, Jun Wang, Robert E. Palermo, and Michael G. Katze. Deep sequencing of HIV-infected cells: insights into nascent transcription and host-directed therapy. *J Virol*, 88(16):8768–8782, Aug 2014. doi: 10.1128/JVI.00768-14. URL http://dx.doi.org/10.1128/JVI.00768-14.

[132] Dinushka Dowling, Somayeh Nasr-Esfahani, Chun H Tan, Kate O'Brien, Jane L Howard, David A Jans, Damian F j Purcell, C. Martin Stoltzfus, and Secondo Sonza. HIV-1 infection induces changes in expression of cellular splicing factors that regulate alternative viral splicing and virus production in macrophages. *Retrovirology*, 5:18, 2008. doi: 10.1186/1742-4690-5-18. URL http://dx.doi.org/10.1186/1742-4690-5-18.

[133] Cynthia de la Fuente, Francisco Santiago, Longwen Deng, Carolyne Eadie, Irene Zilberman, Kylene Kehn, Anil Maddukuri, Shanese Baylor, Kaili Wu, Chee Gun Lee, Anne Pumfery, and Fatah Kashanchi. Gene expression profile of HIV-1 Tat expressing cells: a close interplay between proliferative and differentiation signals. *BMC Biochem*, 3:14, Jun 2002. doi: 10.1186/1471-2091-3-14. URL http://dx.doi.org/10.1186/1471-2091-3-14.

[134] R. Collman, J. W. Balliet, S. A. Gregory, H. Friedman, D. L. Kolson, N. Nathanson, and A. Srinivasan. An infectious molecular clone of an unusual macrophage-tropic and highly cytopathic strain of human immunodeficiency virus type 1. *J Virol*, 66(12):7517–7521, Dec 1992. URL http://jvi.asm.org/content/66/12/7517.

[135] Charles C Berry, Karen Ocwieja, Nirav Malani, and Frederic D Bushman. Comparing DNA integration site clusters with scan statistics. *Bioinformatics*, 30:1493–1500, 2014. doi: 10.1093/bioinformatics/btu035. URL http://bioinformatics.oxfordjournals.org/content/early/2014/01/30/bioinformatics.btu035.abstract.

[136] W. James Kent. BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4):656–664, Apr 2002. doi: 10.1101/gr.229202. URL http://dx.doi.org/10.1101/gr.229202.

[137] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009. doi: 10.1186/gb-2009-10-3-r25. URL http://genomebiology.com/2009/10/3/R25.

[138] Gregory R Grant, Michael H Farkas, Angel D Pizarro, Nicholas F Lahens, Jonathan Schug, Brian P Brunk, Christian J Stoeckert, John B Hogenesch, and Eric A Pierce. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, 27(18):2518–2528, Sep 2011. doi: 10.1093/bioinformatics/btr427. URL http://dx.doi.org/10.1093/bioinformatics/btr427.

[139] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.

[140] Qingsheng Li, Anthony J Smith, Timothy W Schacker, John V Carlis, Lijie Duan, Cavan S Reilly, and Ashley T Haase. Microarray analysis of lymphatic tissue reveals stage-specific, gene expression signatures in HIV-1 infection. *J Immunol*, 183(3):1975–1982, Aug 2009. doi: 10.4049/jimmunol.0803222. URL http://dx.doi.org/10.4049/jimmunol.0803222.

[141] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005. doi: 10.1073/pnas.0506580102. URL http://dx.doi.org/10.1073/pnas.0506580102.

[142] W James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. The human genome browser at UCSC. *Genome Res*, 12(6):996–1006, Jun 2002. doi: 10.1101/gr.229102. URL http://dx.doi.org/10.1101/gr.229102.

[143] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup . The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009. doi: 10.1093/bioinformatics/btp352.

[144] Ravi P. Subramanian, Julia H. Wildschutte, Crystal Russo, and John M. Coffin. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology*, 8: 90, 2011. doi: 10.1186/1742-4690-8-90. URL http://dx.doi.org/10.1186/1742-4690-8-90.

[145] G. La Mantia, D. Maglione, G. Pengue, A. Di Cristofano, A. Simeone, L. Lanfrancone, and L. Lania. Identification and characterization of novel human endogenous retroviral sequences prefentially expressed in undifferentiated embryonal carcinoma cells. *Nucleic Acids Res*, 19(7):1513–1520, Apr 1991. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC333909/.

[146] G. La Mantia, B. Majello, A. Di Cristofano, M. Strazzullo, G. Minchiotti, and L. Lania. Identification of regulatory elements within the minimal promoter region of the human endogenous ERV9 proviruses: accurate transcription initiation is controlled by an Inr-like element. *Nucleic Acids Res*, 20(16):4129–4136, Aug 1992. doi: 10.1093/nar/20.16.4129.

[147] K. E. Plant, S. J. Routledge, and N. J. Proudfoot. Intergenic transcription in the human beta-globin gene cluster. *Mol Cell Biol*, 21(19):6507–6514, Oct 2001. doi: 10.1128/MCB.21.19.6507-6514.2001.

[148] Jianhua Ling, Wenhu Pi, Roni Bollag, Shan Zeng, Meral Keskintepe, Hatem Saliman, Sanford Krantz, Barry Whitney, and Dorothy Tuan. The solitary long terminal repeats of ERV-9 endogenous retrovirus are conserved during primate evolution and possess enhancer activities in embryonic and hematopoietic cellspi. *J Virol*, 76(5):2410–2423, Mar 2002. doi: 10.1128/jvi.76.5.2410-2423.2002. URL http://jvi.asm.org/content/76/5/2410.long.

[149] Xiuping Yu, Xingguo Zhu, Wenhu Pi, Jianhua Ling, Lan Ko, Yoshihiko Takeda, and Dorothy Tuan. The long terminal repeat (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly of an active LTR enhancer complex NF-Y/MZF1/GATA-2. *J Biol Chem*, 280(42):35184–35194, Oct 2005. doi: 10.1074/jbc.M508138200. URL http://dx.doi.org/10.1074/jbc.M508138200.

[150] Robert C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, Aug 2004. doi: 10.1186/1471-2105-5-113. URL http://dx.doi.org/10.1186/1471-2105-5-113.

[151] Margalida Rotger, Judith Dalmau, Andri Rauch, Paul McLaren, Steven E Bosinger, Raquel Martinez, Netanya G Sandler, Annelys Roque, Julia Liebner, Manuel Battegay, Enos Bernasconi, Patrick Descombes, Itziar Erkizia, Jacques Fellay, Bernard Hirschel, Jose M Mir, Eduard Palou, Matthias Hoffmann, Marta Massanella, Juli Blanco, Matthew Woods, Huldrych F Gnthard, Paul de Bakker, Daniel C Douek, Guido Silvestri, Javier Martinez-Picado, and Amalio Telenti. Comparative transcriptomics of extreme phenotypes of human HIV-1 infection and SIV infection in sooty mangabey and rhesus macaque. *J Clin Invest*, 121(6):2391–2400, Jun 2011. doi: 10.1172/JCI45235. URL http://dx.doi.org/10.1172/JCI45235.

[152] Karin Breuer, Amir K. Foroushani, Matthew R. Laird, Carol Chen, Anastasia Sribnaia, Raymond Lo, Geoffrey L. Winsor, Robert E W. Hancock, Fiona S L. Brinkman, and David J. Lynn. InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. *Nucleic Acids Res*, 41 (Database issue):D1228–D1233, Jan 2013. doi: 10.1093/nar/gks1147. URL http://dx.doi.org/10.1093/nar/gks1147.

[153] Irina Rusinova, Sam Forster, Simon Yu, Anitha Kannan, Marion Masse, Helen Cumming, Ross Chapman, and Paul J. Hertzog. Interferome v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res*, 41(Database issue):D1040–D1046, Jan 2013. doi: 10.1093/nar/gks1215. URL http://dx.doi.org/10.1093/nar/gks1215.

[154] Stewart T. Chang, Matthew J. Thomas, Pavel Sova, Richard R. Green, Robert E. Palermo, and Michael G. Katze. Next-generation sequencing of small RNAs from HIV-infected cells identifies phased microrna expression patterns and candidate novel microRNAs differentially expressed upon infection. *MBio*, 4(1):e00549–

e00512, 2013. doi: 10.1128/mBio.00549-12. URL http://dx.doi.org/10.1128/mBio.00549-12.

[155] Zeynep Kalender Atak, Kim De Keersmaecker, Valentina Gianfelici, Ellen Geerdens, Roel Vandepoel, Daphnie Pauwels, Michaël Porcu, Idoya Lahortiga, Vanessa Brys, Willy G. Dirks, Hilmar Quentmeier, Jacqueline Cloos, Harry Cuppens, Anne Uyttebroeck, Peter Vandenberghe, Jan Cools, and Stein Aerts. High accuracy mutation detection in leukemia on a selected panel of cancer genes. *PLoS One*, 7(6):e38463, 2012. doi: 10.1371/journal.pone.0038463. URL http://dx.doi.org/10.1371/journal.pone.0038463.

[156] Ekta S. Patel and Lung-Ji Chang. Synergistic effects of interleukin-7 and pre-T cell receptor signaling in human T cell development. *J Biol Chem*, 287(40):33826–33835, Sep 2012. doi: 10.1074/jbc.M112.380113. URL http://dx.doi.org/10.1074/jbc.M112.380113.

[157] Michaël Imbeault, Michel Ouellet, and Michel J. Tremblay. Microarray study reveals that HIV-1 induces rapid type-i interferon-dependent p53 mRNA up-regulation in human primary CD4+ T cells. *Retrovirology*, 6:5, 2009. doi: 10.1186/1742-4690-6-5. URL http://dx.doi.org/10.1186/1742-4690-6-5.

[158] S. Iwase, Y. Furukawa, J. Kikuchi, M. Nagai, Y. Terui, M. Nakamura, and H. Yamada. Modulation of E2F activity is linked to interferon-induced growth suppression of hematopoietic cells. *J Biol Chem*, 272(19):12406–12414, May 1997. doi: 10.1074/jbc.272.19.12406. URL http://dx.doi.org/10.1074/jbc.272.19.12406.

[159] R. W. Johnstone, J. A. Kerry, and J. A. Trapani. The human interferon-inducible protein, IFI 16, is a repressor of transcription. *J Biol Chem*, 273(27):17172–17177, Jul 1998. doi: 10.1074/jbc.273.27.17172. URL http://dx.doi.org/10.1074/jbc.273.27.17172.

[160] B. R. Williams. PKR; a sentinel kinase for cellular stress. *Oncogene*, 18(45):6112–6120, Nov 1999. doi: 10.1038/sj.onc.1203127. URL http://dx.doi.org/10.1038/sj.onc.1203127.

[161] C. V. Ramana, N. Grammatikakis, M. Chernov, H. Nguyen, K. C. Goh, B. R. Williams, and G. R. Stark. Regulation of c-myc expression by IFN-gamma through Stat1-dependent and -independent pathways. *EMBO J*, 19(2):263–272, Jan 2000. doi: 10.1093/emboj/19.2.263. URL http://dx.doi.org/10.1093/emboj/19.2.263.

[162] Shu-Ling Liang, David Quirk, and Aimin Zhou. RNase L: its biological roles and regulation. *IUBMB Life*, 58(9):508–514, Sep 2006. doi: 10.1080/15216540600838232. URL http://dx.doi.org/10.1080/15216540600838232.

[163] Fan Hsu, W. James Kent, Hiram Clawson, Robert M Kuhn, Mark Diekhans, and David Haussler. The UCSC known genes. *Bioinformatics*, 22(9):1036–1046, May 2006. doi: 10.1093/bioinformatics/btl048. URL http://dx.doi.org/10.1093/bioinformatics/btl048.

[164] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, 110(1-4):462–467, 2005. doi: 10.1159/000084979. URL http://dx.doi.org/10.1159/000084979.

[165] F. Maldarelli, C. Xiang, G. Chamoun, and S. L. Zeichner. The expression of the essential nuclear splicing factor SC35 is altered by human immunodeficiency virus infection. *Virus Res*, 53(1):39–51, Jan 1998. URL http://www.sciencedirect.com/science/article/pii/S0168170297001305.

[166] Anne Monette, Lara Ajamian, Marcelo López-Lastra, and Andrew J. Mouland. Human immunodeficiency virus type 1 (HIV-1) induces the cytoplasmic retention of heterogeneous nuclear ribonucleoprotein A1 by disrupting nuclear import: implications for HIV-1 gene expression. *J Biol Chem*, 284(45):31350–31362, Nov 2009. doi: 10.1074/jbc.M109.048736. URL http://dx.doi.org/10.1074/jbc.M109.048736.

[167] Rafael Contreras-Galindo, Pablo López, Rosa Vélez, and Yasuhiro Yamamura. HIV-1 infection increases the expression of human endogenous retroviruses type K (HERV-K) in vitro. *AIDS Res Hum Retroviruses*, 23(1):116–122, Jan 2007. doi: 10.1089/aid.2006.0117. URL http://dx.doi.org/10.1089/aid.2006.0117.

[168] Rafael Contreras-Galindo, Mark H Kaplan, Shirley He, Angie C Contreras-Galindo, Marta J Gonzalez-Hernandez, Ferdinand Kappes, Derek Dube, Susana M Chan, Dan Robinson, Fan Meng, Manhong Dai, Scott D Gitlin, Arul M Chinnaiyan, Gilbert S Omenn, and David M Markovitz. HIV infection reveals widespread expansion of novel centromeric human endogenous retroviruses. *Genome Res*, 23(9):1505–1513, Sep 2013. doi: 10.1101/gr.144303.112. URL http://dx.doi.org/10.1101/gr.144303.112.

[169] Neeru Bhardwaj, Frank Maldarelli, John Mellors, and John M. Coffin. HIV-1 infection leads to increased transcription of human endogenous retrovirus HERV-K (HML-2) proviruses in vivo but not to increased virion production. *J Virol*, 88(19):11108–11120, Oct 2014. doi: 10.1128/JVI.01623-14. URL http://dx.doi.org/10.1128/JVI.01623-14.

[170] R Brad Jones, Haihan Song, Yang Xu, Keith E. Garrison, Anton A. Buzdin, Naveed Anwar, Diana V. Hunter, Shariq Mujib, Vesna Mihajlovic, Eric Martin, Erika Lee, Monika Kuciak, Rui André Saraiva Raposo, Ardalan Bozorgzad, Duncan A. Meiklejohn, Lishomwa C. Ndhlovu, Douglas F. Nixon, and Mario A. Ostrowski. LINE-1 retrotransposable element DNA accumulates in HIV-1-infected

cells. *J Virol*, 87(24):13307–13320, Dec 2013. doi: 10.1128/JVI.02257-13. URL http://dx.doi.org/10.1128/JVI.02257-13.

[171] P. Medstrand and D. L. Mager. Human-specific integrations of the HERV-K endogenous retrovirus family. *J Virol*, 72(12):9782–9787, Dec 1998. URL jvi.asm.org/cgi/pmidlookup?view=long&pmid=9811713.

[172] Catriona Macfarlane and Peter Simmonds. Allelic variation of HERV-K(HML-2) endogenous retroviral elements in human populations. *J Mol Evol*, 59(5):642–656, Nov 2004. doi: 10.1007/s00239-004-2656-1. URL http://dx.doi.org/10.1007/s00239-004-2656-1.

[173] Kristina Büscher, Uwe Trefzer, Maja Hofmann, Wolfram Sterry, Reinhard Kurth, and Joachim Denner. Expression of human endogenous retrovirus K in melanomas and melanoma cell lines. *Cancer Res*, 65(10):4172–4180, May 2005. doi: 10.1158/0008-5472.CAN-04-2983. URL http://dx.doi.org/10.1158/0008-5472.CAN-04-2983.

[174] G. Howard, R. Eiges, F. Gaudet, R. Jaenisch, and A. Eden. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. *Oncogene*, 27(3):404–408, Jan 2008. doi: 10.1038/sj.onc.1210631. URL http://dx.doi.org/10.1038/sj.onc.1210631.

[175] Rebecca C. Iskow, Michael T. McCabe, Ryan E. Mills, Spencer Torene, W Stephen Pittard, Andrew F. Neuwald, Erwin G. Van Meir, Paula M. Vertino, and Scott E. Devine. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, 141(7):1253–1261, Jun 2010. doi: 10.1016/j.cell.2010.05.020. URL http://dx.doi.org/10.1016/j.cell.2010.05.020.

[176] Eunjung Lee, Rebecca Iskow, Lixing Yang, Omer Gokcumen, Psalm Haseley, Lovelace J Luquette, 3rd, Jens G. Lohr, Christopher C. Harris, Li Ding, Richard K. Wilson, David A. Wheeler, Richard A. Gibbs, Raju Kucherlapati, Charles Lee, Peter V. Kharchenko, Peter J. Park, and Cancer Genome Atlas Research Network . Landscape of somatic retrotransposition in human cancers. *Science*, 337(6097):967–971, Aug 2012. doi: 10.1126/science.1222077. URL http://dx.doi.org/10.1126/science.1222077.

[177] Steven W. Criscione, Yue Zhang, William Thompson, John M. Sedivy, and Nicola Neretti. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics*, 15:583, 2014. doi: 10.1186/1471-2164-15-583. URL http://dx.doi.org/10.1186/1471-2164-15-583.

[178] Adam W Whisnant, Hal P Bogerd, Omar Flores, Phong Ho, Jason G Powers, Natalia Sharova, Mario Stevenson, Chin-Ho Chen, and Bryan R Cullen. In-depth analysis of the interaction of HIV-1 with cellular microRNA biogenesis and effector mechanisms. *MBio*, 4(2):e000193, 2013. doi: 10.1128/mBio.00193-13. URL http://dx.doi.org/10.1128/mBio.00193-13.

[179] Nicholas F. Lahens, Ibrahim Halil Kavakli, Ray Zhang, Katharina Hayer, Michael B. Black, Hannah Dueck, Angel Pizarro, Junhyong Kim, Rafael Irizarry, Russell S. Thomas, Gregory R. Grant, and John B. Hogenesch. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol*, 15(6):R86, 2014. doi: 10.1186/gb-2014-15-6-r86. URL http://dx.doi.org/10.1186/gb-2014-15-6-r86.

[180] R. D. Hockett, J. M. Kilby, C. A. Derdeyn, M. S. Saag, M. Sillers, K. Squires, S. Chiz, M. A. Nowak, G. M. Shaw, and R. P. Bucy. Constant mean viral copy number per infected cell in tissues regardless of high, low, or undetectable plasma HIV RNA. *J Exp Med*, 189(10):1545–1554, May 1999. doi: 10.1084/jem.189.10.1545. URL http://jem.rupress.org/content/189/10/1545.long.

[181] Rob J. De Boer, Ruy M. Ribeiro, and Alan S. Perelson. Current estimates for HIV-1 production imply rapid viral clearance in lymphoid tissues. *PLoS Comput Biol*, 6(9):e1000906, 2010. doi: 10.1371/journal.pcbi.1000906. URL http://dx.doi.org/10.1371/journal.pcbi.1000906.

[182] Terumasa Ikeda, Junji Shibata, Kazuhisa Yoshimura, Atsushi Koito, and Shuzo Matsushita. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. *J Infect Dis*, 195(5):716–725, Mar 2007. doi: 10.1086/510915. URL http://dx.doi.org/10.1086/510915.

[183] Thor A. Wagner, Sherry McLaughlin, Kavita Garg, Charles Y K. Cheung, Brendan B. Larsen, Sheila Styrchak, Hannah C. Huang, Paul T. Edlefsen, James I. Mullins, and Lisa M. Frenkel. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science*, Jul 2014. doi: 10.1126/science.1256304. URL http://dx.doi.org/10.1126/science.1256304.

[184] F. Maldarelli, X. Wu, L. Su, F. R. Simonetti, W. Shao, S. Hill, J. Spindler, A. L. Ferris, J. W. Mellors, M. F. Kearney, J. M. Coffin, and S. H. Hughes. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science*, 345:179–183, Jun 2014. doi: 10.1126/science.1254194. URL http://dx.doi.org/10.1126/science.1254194.

[185] Lillian B. Cohn, Israel T. Silva, Thiago Y. Oliveira, Rafael A. Rosales, Erica H. Parrish, Gerald H. Learn, Beatrice H. Hahn, Julie L. Czartoski, M Juliana McElrath, Clara Lehmann, Florian Klein, Marina Caskey, Bruce D. Walker, Janet D. Siliciano, Robert F. Siliciano, Mila Jankovic, and Michel C. Nussenzweig. HIV-1 integration landscape during latent and active infection. *Cell*, 160(3):420–432, Jan 2015. doi: 10.1016/j.cell.2015.01.020. URL http://dx.doi.org/10.1016/j.cell.2015.01.020.

[186] Astrid R W Schröder, Paul Shinn, Huaming Chen, Charles Berry, Joseph R Ecker, and Frederic Bushman. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, 110(4):521–529, Aug 2002.

doi: 10.1016/S0092-8674(02)00864-4. URL http://dx.doi.org/10.1016/S0092-8674(02)00864-4.

[187] Chunlin Wang, Yumi Mitsuya, Baback Gharizadeh, Mostafa Ronaghi, and Shafer. Robert W. Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance. *Genome Research*, 17:1195–1201, 2007. doi: 10.1101/gr.6468307. URL http://genome.cshlp.org/content/17/8/1195.

[188] Troy Brady, Young Nam Lee, Keshet Ronen, Nirav Malani, Charles C Berry, Paul D Bieniasz, and Frederic D Bushman. Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev*, 23(5):633–642, Mar 2009. doi: 10.1101/gad.1762309. URL http://dx.doi.org/10.1101/gad.1762309.

[189] Scott Sherrill-Mix, Mary K. Lewinski, Marylinda Famiglietti, Alberto Bosque, Nirav Malani, Karen E. Ocwieja, Charles C. Berry, David Looney, Liang Shan, Luis M. Agosto, Matthew J. Pace, Robert F. Siliciano, Una O'Doherty, John Guatelli, Vicente Planelles, and Frederic D. Bushman. HIV latency and integration site placement in five cell-based models. *Retrovirology*, 10:90, 2013. doi: 10.1186/1742-4690-10-90. URL http://dx.doi.org/10.1186/1742-4690-10-90.

[190] Bruna Marini, Attila Kertesz-Farkas, Hashim Ali, Bojana Lucic, Kamil Lisek, Lara Manganaro, Sandor Pongor, Roberto Luzzati, Alessandra Recchia, Fulvio Mavilio, Mauro Giacca, and Marina Lusic. Nuclear architecture dictates HIV-1 integration site selection. *Nature*, Mar 2015. doi: 10.1038/nature14226. URL http://dx.doi.org/10.1038/nature14226.

[191] Marina Cavazzana-Calvo, Emmanuel Payen, Olivier Negre, Gary Wang, Kathleen Hehir, Floriane Fusil, Julian Down, Maria Denaro, Troy Brady, Karen Westerman, Resy Cavallesco, Beatrix Gillet-Legrand, Laure Caccavelli, Riccardo Sgarra, Leila Maouche-Chrétien, Françoise Bernaudin, Robert Girot, Ronald Dorazio, Geert-Jan Mulder, Axel Polack, Arthur Bank, Jean Soulier, Jérôme Larghero, Nabil Kabbara, Bruno Dalle, Bernard Gourmel, Gérard Socie, Stany Chrétien, Nathalie Cartier, Patrick Aubourg, Alain Fischer, Kenneth Cornetta, Frédéric Galacteros, Yves Beuzard, Eliane Gluckman, Frederick Bushman, Salima Hacein-Bey-Abina, and Philippe Leboulch. Transfusion independence and HMGA2 activation after gene therapy of human β-thalassaemia. *Nature*, 467(7313):318–322, Sep 2010. doi: 10.1038/nature09328. URL http://dx.doi.org/10.1038/nature09328.

[192] Salima Hacein-Bey-Abina, Alexandrine Garrigue, Gary P. Wang, Jean Soulier, Annick Lim, Estelle Morillon, Emmanuelle Clappier, Laure Caccavelli, Eric Delabesse, Kheira Beldjord, Vahid Asnafi, Elizabeth MacIntyre, Liliane Dal Cortivo, Isabelle Radford, Nicole Brousse, François Sigaux, Despina Moshous,

Julia Hauer, Arndt Borkhardt, Bernd H. Belohradsky, Uwe Wintergerst, Maria C. Velez, Lily Leiva, Ricardo Sorensen, Nicolas Wulffraat, Stéphane Blanche, Frederic D. Bushman, Alain Fischer, and Marina Cavazzana-Calvo. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J Clin Invest*, 118(9):3132–3142, Sep 2008. doi: 10.1172/JCI35700. URL http://dx.doi.org/10.1172/JCI35700.

[193] Arianna Moiani, Ylenia Paleari, Daniela Sartori, Riccardo Mezzadra, Annarita Miccio, Claudia Cattoglio, Fabienne Cocchiarella, Maria Rosa Lidonnici, Giuliana Ferrari, and Fulvio Mavilio. Lentiviral vector integration in the human genome induces alternative splicing and generates aberrant transcripts. *J Clin Invest*, 122(5):1653–1666, May 2012. doi: 10.1172/JCI61852. URL http://dx.doi.org/10.1172/JCI61852.

[194] Daniela Cesana, Jacopo Sgualdino, Laura Rudilosso, Stefania Merella, Luigi Naldini, and Eugenio Montini. Whole transcriptome characterization of aberrant splicing events induced by lentiviral vector integrations. *J Clin Invest*, 122(5):1667–1676, May 2012. doi: 10.1172/JCI62189. URL http://dx.doi.org/10.1172/JCI62189.

[195] S. Pääbo, D. M. Irwin, and A. C. Wilson. DNA damage promotes jumping between templates during enzymatic amplification. *J Biol Chem*, 265(8):4718–4721, Mar 1990. URL http://www.jbc.org/content/265/8/4718.

[196] S. J. Odelberg, R. B. Weiss, A. Hata, and R. White. Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. *Nucleic Acids Res*, 23(11):2049–2057, Jun 1995. doi: 10.1093/nar/23.11.2049. URL http://dx.doi.org/10.1093/nar/23.11.2049.

[197] Xian-Chun Zeng and San-Xia Wang. Evidence that BmTXK beta-BmKCT cDNA from Chinese scorpion Buthus martensii Karsch is an artifact generated in the reverse transcription process. *FEBS Lett*, 520(1-3):183–4; author reply 185, Jun 2002. URL http://www.sciencedirect.com/science/article/pii/S0014579302028120.

[198] Bosiljka Tasic, Christoph E. Nabholz, Kristin K. Baldwin, Youngwook Kim, Erroll H. Rueckert, Scott A. Ribich, Paula Cramer, Qiang Wu, Richard Axel, and Tom Maniatis. Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. *Mol Cell*, 10(1):21–33, Jul 2002. URL http://www.sciencedirect.com/science/article/pii/S1097276502005786.

[199] Miklós Geiszt, Kristen Lekstrom, and Thomas L. Leto. Analysis of mRNA transcripts from the NAD(P)H oxidase 1 (Nox1) gene. evidence against production of the NADPH oxidase homolog-1 short (NOH-1S) transcript variant. *J Biol Chem*, 279(49):51661–51668, Dec 2004. doi: 10.1074/jbc.M409325200. URL http://dx.doi.org/10.1074/jbc.M409325200.

[200] Julie Cocquet, Allen Chong, Guanglan Zhang, and Reiner A. Veitia. Reverse transcriptase template switching and false alternative transcripts. *Genomics*, 88 (1):127–131, Jul 2006. doi: 10.1016/j.ygeno.2005.12.013. URL http://dx.doi.org/10.1016/j.ygeno.2005.12.013.

[201] C. Joel McManus, Joseph D Coolon, Michael O Duff, Jodi Eipper-Mains, Brenton R Graveley, and Patricia J Wittkopp. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res*, 20(6):816–825, Jun 2010. doi: 10.1101/gr.102491.109. URL http://dx.doi.org/10.1101/gr.102491.109.

[202] Benjamin Cogné, Richard Snyder, Pierre Lindenbaum, Jean-Baptiste Dupont, Richard Redon, Philippe Moullier, and Adrien Leger. NGS library preparation may generate artifactual integration sites of AAV vectors. *Nat Med*, 20(6):577–578, Jun 2014. doi: 10.1038/nm.3578. URL http://dx.doi.org/10.1038/nm.3578.

[203] E. Gilboa, S. W. Mitra, S. Goff, and D. Baltimore. A detailed model of reverse transcription and tests of crucial aspects. *Cell*, 18(1):93–100, Sep 1979. doi: 10.1016/0092-8674(79)90357-X. URL http://dx.doi.org/10.1016/0092-8674(79)90357-X.

[204] G. X. Luo and J. Taylor. Template switching by reverse transcriptase during DNA synthesis. *J Virol*, 64(9):4321–4328, Sep 1990. URL http://jvi.asm.org/content/64/9/4321.long.

[205] Jonathan Houseley and David Tollervey. Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS One*, 5(8):e12271, 2010. doi: 10.1371/journal.pone.0012271. URL http://dx.doi.org/10.1371/journal.pone.0012271.

[206] A. Meyerhans, J. P. Vartanian, and S. Wain-Hobson. DNA recombination during PCR. *Nucleic Acids Res*, 18(7):1687–1691, Apr 1990. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC330584/.

[207] Daniel J G. Lahr and Laura A. Katz. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques*, 47(4):857–866, Oct 2009. doi: 10.2144/000113219. URL http://dx.doi.org/10.2144/000113219.

[208] W. Al-Ahmadi, L. Al-Haj, F. A. Al-Mohanna, R. H. Silverman, and K S A. Khabar. RNase L downmodulation of the RNA-binding protein, HuR, and cellular growth. *Oncogene*, 28(15):1782–1791, Apr 2009. doi: 10.1038/onc.2009.16. URL http://dx.doi.org/10.1038/onc.2009.16.

[209] R Brad Jones, Keith E. Garrison, Shariq Mujib, Vesna Mihajlovic, Nasra Aidarus, Diana V. Hunter, Eric Martin, Vivek M. John, Wei Zhan, Nabil F. Faruk, Gabor Gyenes, Neil C. Sheppard, Ingrid M. Priumboom-Brees, David A. Goodwin,

Lianchun Chen, Melanie Rieger, Sophie Muscat-King, Peter T. Loudon, Cole Stanley, Sara J. Holditch, Jessica C. Wong, Kiera Clayton, Erick Duan, Haihan Song, Yang Xu, Devi SenGupta, Ravi Tandon, Jonah B. Sacha, Mark A. Brockman, Erika Benko, Colin Kovacs, Douglas F. Nixon, and Mario A. Ostrowski. HERV-K-specific T cells eliminate diverse HIV-1/2 and SIV primary isolates. *J Clin Invest*, 122(12):4473–4489, Dec 2012. doi: 10.1172/JCI64560. URL http://dx.doi.org/10.1172/JCI64560.

[210] K. Boller, O. Janssen, H. Schuldes, R. R. Tönjes, and R. Kurth. Characterization of the antibody response specific for the human endogenous retrovirus HTDV/HERV-K. *J Virol*, 71(6):4581–4588, Jun 1997. URL http://jvi.asm.org/content/71/6/4581.

[211] Keith E. Garrison, R Brad Jones, Duncan A. Meiklejohn, Naveed Anwar, Lishomwa C. Ndhlovu, Joan M. Chapman, Ann L. Erickson, Ashish Agrawal, Gerald Spotts, Frederick M. Hecht, Seth Rakoff-Nahoum, Jack Lenz, Mario A. Ostrowski, and Douglas F. Nixon. T cell responses to human endogenous retroviruses in HIV-1 infection. *PLoS Pathog*, 3(11):e165, Nov 2007. doi: 10.1371/journal.ppat.0030165. URL http://dx.doi.org/10.1371/journal.ppat.0030165.

[212] Ravi Tandon, Devi SenGupta, Lishomwa C. Ndhlovu, Raphaella G S. Vieira, R Brad Jones, Vanessa A. York, Vinicius A. Vieira, Elizabeth R. Sharp, Andrew A. Wiznia, Mario A. Ostrowski, Michael G. Rosenberg, and Douglas F. Nixon. Identification of human endogenous retrovirus-specific T cell responses in vertically HIV-1-infected subjects. *J Virol*, 85(21):11526–11531, Nov 2011. doi: 10.1128/JVI.05418-11. URL http://dx.doi.org/10.1128/JVI.05418-11.

[213] Devi SenGupta, Ravi Tandon, Raphaella G S. Vieira, Lishomwa C. Ndhlovu, Rachel Lown-Hecht, Christopher E. Ormsby, Liyen Loh, R Brad Jones, Keith E. Garrison, Jeffrey N. Martin, Vanessa A. York, Gerald Spotts, Gustavo Reyes-Terán, Mario A. Ostrowski, Frederick M. Hecht, Steven G. Deeks, and Douglas F. Nixon. Strong human endogenous retrovirus-specific T cell responses are associated with control of HIV-1 in chronic infection. *J Virol*, 85(14):6977–6985, Jul 2011. doi: 10.1128/JVI.00179-11. URL http://dx.doi.org/10.1128/JVI.00179-11.

[214] Wenhu Pi, Zhongan Yang, Jian Wang, Ling Ruan, Xiuping Yu, Jianhua Ling, Sanford Krantz, Carlos Isales, Simon J. Conway, Shuo Lin, and Dorothy Tuan. The LTR enhancer of ERV-9 human endogenous retrovirus is active in oocytes and progenitor cells in transgenic zebrafish and humans. *Proc Natl Acad Sci U S A*, 101(3):805–810, Jan 2004. doi: 10.1073/pnas.0307698100. URL http://dx.doi.org/10.1073/pnas.0307698100.

[215] Xiang H-F Zhang and Lawrence A Chasin. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc Natl Acad Sci*

USA, 103(36):13427–13432, Sep 2006. doi: 10.1073/pnas.0603042103. URL http://dx.doi.org/10.1073/pnas.0603042103.

[216] J. Smith, A. Azad, and N. Deacon. Identification of two novel human immunodeficiency virus type 1 splice acceptor sites in infected T cell lines. *J Gen Virol*, 73 (Pt 7):1825–1828, Jul 1992. URL http://vir.sgmjournals.org/content/73/7/1825.long.

[217] Cristina Carrera, Milagros Pinilla, Lucía Pérez-Alvarez, and Michael M. Thomson. Identification of unusual and novel HIV type 1 spliced transcripts generated in vivo. *AIDS Res Hum Retroviruses*, 26(7):815–820, Jul 2010. doi: 10.1089/aid.2010.0011. URL http://dx.doi.org/10.1089/aid.2010.0011.

[218] Federico A. Santoni, Jessica Guerra, and Jeremy Luban. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology*, 9:111, 2012. doi: 10.1186/1742-4690-9-111. URL http://dx.doi.org/10.1186/1742-4690-9-111.

[219] Nina V. Fuchs, Sabine Loewer, George Q. Daley, Zsuzsanna Izsvák, Johannes Löwer, and Roswitha Löwer. Human endogenous retrovirus K (HML-2) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. *Retrovirology*, 10:115, 2013. doi: 10.1186/1742-4690-10-115. URL http://dx.doi.org/10.1186/1742-4690-10-115.

[220] Alexandre Fort, Kosuke Hashimoto, Daisuke Yamada, Md Salimullah, Chaman A. Keya, Alka Saxena, Alessandro Bonetti, Irina Voineagu, Nicolas Bertin, Anton Kratz, Yukihiko Noro, Chee-Hong Wong, Michiel de Hoon, Robin Andersson, Albin Sandelin, Harukazu Suzuki, Chia-Lin Wei, Haruhiko Koseki, F. A. N. T. O. M Consortium , Yuki Hasegawa, Alistair R R. Forrest, and Piero Carninci. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat Genet*, 46(6):558–566, Jun 2014. doi: 10.1038/ng.2965. URL http://dx.doi.org/10.1038/ng.2965.

[221] Jichang Wang, Gangcai Xie, Manvendra Singh, Avazeh T. Ghanbarian, Tamás Raskó, Attila Szvetnik, Huiqiang Cai, Daniel Besser, Alessandro Prigione, Nina V. Fuchs, Gerald G. Schumann, Wei Chen, Matthew C. Lorincz, Zoltán Ivics, Laurence D. Hurst, and Zsuzsanna Izsvák. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, 516(7531):405–409, Dec 2014. doi: 10.1038/nature13804. URL http://dx.doi.org/10.1038/nature13804.

[222] Beda Joos, Marek Fischer, Herbert Kuster, Satish K. Pillai, Joseph K. Wong, Jürg Böni, Bernard Hirschel, Rainer Weber, Alexandra Trkola, Huldrych F. Günthard, and Swiss H. I. V Cohort Study . HIV rebounds from latently infected cells, rather than from continuing low-level replication. *Proc Natl Acad Sci U S A*, 105(43):16725–16730, Oct 2008. doi: 10.1073/pnas.0804192105. URL http://dx.doi.org/10.1073/pnas.0804192105.

[223] Timothy P. Brennan, John O. Woods, Ahmad R. Sedaghat, Janet D. Siliciano, Robert F. Siliciano, and Claus O. Wilke. Analysis of human immunodeficiency virus type 1 viremia and provirus in resting CD4+ T cells reveals a novel source of residual viremia in patients on antiretroviral therapy. *J Virol*, 83(17):8470–8481, Sep 2009. doi: 10.1128/JVI.02568-08. URL http://dx.doi.org/10.1128/JVI.02568-08.

[224] Thor A. Wagner, Jen L. McKernan, Nicole H. Tobin, Ken A. Tapia, James I. Mullins, and Lisa M. Frenkel. An increasing proportion of monotypic HIV-1 DNA sequences during antiretroviral treatment suggests proliferation of HIV-infected cells. *J Virol*, 87(3):1770–1778, Feb 2013. doi: 10.1128/JVI.01985-12. URL http://dx.doi.org/10.1128/JVI.01985-12.

[225] Mary F. Kearney, Jonathan Spindler, Wei Shao, Sloane Yu, Elizabeth M. Anderson, Angeline O'Shea, Catherine Rehm, Carry Poethke, Nicholas Kovacs, John W. Mellors, John M. Coffin, and Frank Maldarelli. Lack of detectable HIV-1 molecular evolution during suppressive antiretroviral therapy. *PLoS Pathog*, 10(3):e1004010, Mar 2014. doi: 10.1371/journal.ppat.1004010. URL http://dx.doi.org/10.1371/journal.ppat.1004010.

[226] T. W. Chun, D. Finzi, J. Margolick, K. Chadwick, D. Schwartz, and R. F. Siliciano. In vivo fate of HIV-1-infected T cells: quantitative analysis of the transition to stable latency. *Nat Med*, 1(12):1284–1290, Dec 1995. URL http://www.nature.com/nm/journal/v1/n12/full/nm1295-1284.html.

[227] T. W. Chun, L. Carruth, D. Finzi, X. Shen, J. A. DiGiuseppe, H. Taylor, M. Hermankova, K. Chadwick, J. Margolick, T. C. Quinn, Y. H. Kuo, R. Brookmeyer, M. A. Zeiger, P. Barditch-Crovo, and R. F. Siliciano. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature*, 387(6629):183–188, May 1997. doi: 10.1038/387183a0. URL http://dx.doi.org/10.1038/387183a0.

[228] R. T. Davey, N. Bhat, C. Yoder, T. W. Chun, J. A. Metcalf, R. Dewar, V. Natarajan, R. A. Lempicki, J. W. Adelsberger, K. D. Miller, J. A. Kovacs, M. A. Polis, R. E. Walker, J. Falloon, H. Masur, D. Gee, M. Baseler, D. S. Dimitrov, A. S. Fauci, and H. C. Lane. HIV-1 and T cell dynamics after interruption of highly active antiretroviral therapy (HAART) in patients with a history of sustained viral suppression. *Proc Natl Acad Sci U S A*, 96(26):15109–15114, Dec 1999. URL http://dx.doi.org/10.1073/pnas.96.26.15109.

[229] Douglas D Richman, David M Margolis, Martin Delaney, Warner C Greene, Daria Hazuda, and Roger J Pomerantz. The challenge of finding a cure for HIV infection. *Science*, 323(5919):1304–1307, Mar 2009. doi: 10.1126/science.1165706. URL http://dx.doi.org/10.1126/science.1165706.

[230] D. Finzi, J. Blankson, J. D. Siliciano, J. B. Margolick, K. Chadwick, T. Pierson, K. Smith, J. Lisziewicz, F. Lori, C. Flexner, T. C. Quinn, R. E. Chaisson, E. Rosenberg, B. Walker, S. Gange, J. Gallant, and R. F. Siliciano. Latent infection of

CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat Med*, 5(5):512–517, May 1999. doi: 10.1038/8394. URL http://dx.doi.org/10.1038/8394.

[231] Janet D Siliciano, Joleen Kajdas, Diana Finzi, Thomas C Quinn, Karen Chadwick, Joseph B Margolick, Colin Kovacs, Stephen J Gange, and Robert F Siliciano. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat Med*, 9(6):727–728, Jun 2003. doi: 10.1038/nm880. URL http://dx.doi.org/10.1038/nm880.

[232] D. Finzi, M. Hermankova, T. Pierson, L. M. Carruth, C. Buck, R. E. Chaisson, T. C. Quinn, K. Chadwick, J. Margolick, R. Brookmeyer, J. Gallant, M. Markowitz, D. D. Ho, D. D. Richman, and R. F. Siliciano. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science*, 278(5341): 1295–1300, Nov 1997. doi: 10.1126/science.278.5341.1295. URL http://dx.doi.org/10.1126/science.278.5341.1295.

[233] Leor S Weinberger, Roy D Dar, and Michael L Simpson. Transient-mediated fate determination in a transcriptional circuit of HIV. *Nat Genet*, 40(4):466–470, Apr 2008. doi: 10.1038/ng.116. URL http://dx.doi.org/10.1038/ng.116.

[234] Abhyudai Singh, Brandon Razooky, Chris D Cox, Michael L Simpson, and Leor S Weinberger. Transcriptional bursting from the HIV-1 promoter is a significant source of stochastic noise in HIV-1 gene expression. *Biophys J*, 98(8):L32–L34, Apr 2010. doi: 10.1016/j.bpj.2010.03.001. URL http://dx.doi.org/10.1016/j.bpj.2010.03.001.

[235] Brandon S Razooky and Leor S Weinberger. Mapping the architecture of the HIV-1 Tat circuit: A decision-making circuit that lacks bistability and exploits stochastic noise. *Methods*, 53(1):68–77, Jan 2011. doi: 10.1016/j.ymeth.2010.12.006. URL http://dx.doi.org/10.1016/j.ymeth.2010.12.006.

[236] H. J. Muller. Types of visible variations induced by X-rays in *Drosophila*. *J Genet*, 22:299–334, 1930. URL http://www.ias.ac.in/j_archive/jgenet/22/vol22contents.html.

[237] Miklos Gaszner and Gary Felsenfeld. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet*, 7(9):703–713, Sep 2006. doi: 10.1038/nrg1925. URL http://dx.doi.org/10.1038/nrg1925.

[238] A. Jordan, P. Defechereux, and E. Verdin. The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *EMBO J*, 20(7):1726–1738, Apr 2001. doi: 10.1093/emboj/20.7.1726. URL http://dx.doi.org/10.1093/emboj/20.7.1726.

[239] Albert Jordan, Dwayne Bisgrove, and Eric Verdin. HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. *EMBO J*, 22(8):1868–1877,

Apr 2003. doi: 10.1093/emboj/cdg188. URL http://dx.doi.org/10.1093/emboj/cdg188.

[240] Richard Pearson, Young Kyeung Kim, Joseph Hokello, Kara Lassen, Julia Friedman, Mudit Tyagi, and Jonathan Karn. Epigenetic silencing of human immunodeficiency virus (HIV) transcription by formation of restrictive chromatin structures at the viral long terminal repeat drives the progressive entry of HIV into latency. *J Virol*, 82(24):12291–12303, Dec 2008. doi: 10.1128/JVI.01383-08. URL http://dx.doi.org/10.1128/JVI.01383-08.

[241] F. Romerio, M. N. Gabriel, and D. M. Margolis. Repression of human immunodeficiency virus type 1 through the novel cooperation of human factors YY1 and LSF. *J Virol*, 71(12):9375–9382, Dec 1997. URL http://jvi.asm.org/content/71/12/9375.long.

[242] J. J. Coull, F. Romerio, J. M. Sun, J. L. Volker, K. M. Galvin, J. R. Davie, Y. Shi, U. Hansen, and D. M. Margolis. The human factors YY1 and LSF repress the human immunodeficiency virus type 1 long terminal repeat via recruitment of histone deacetylase 1. *J Virol*, 74(15):6790–6799, Aug 2000. doi: 10.1128/JVI.74.15.6790-6799.2000. URL http://jvi.asm.org/content/74/15/6790.

[243] Guocheng He and David M Margolis. Counterregulation of chromatin deacetylation and histone deacetylase occupancy at the integrated promoter of human immunodeficiency virus type 1 (HIV-1) by the HIV-1 repressor YY1 and HIV-1 activator Tat. *Mol Cell Biol*, 22(9):2965–2973, May 2002. doi: 10.1128/MCB.22.9.2965-2973.2002. URL http://dx.doi.org/10.1128/MCB.22.9.2965-2973.2002.

[244] M. K. Lewinski, D. Bisgrove, P. Shinn, H. Chen, C. Hoffmann, S. Hannenhalli, E. Verdin, C. C. Berry, J. R. Ecker, and F. D. Bushman. Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J Virol*, 79(11):6610–6619, Jun 2005. doi: 10.1128/JVI.79.11.6610-6619.2005. URL http://dx.doi.org/10.1128/JVI.79.11.6610-6619.2005.

[245] Liang Shan, Hung-Chih Yang, S. Alireza Rabi, Hector C Bravo, Neeta S Shroff, Rafael A Irizarry, Hao Zhang, Joseph B Margolick, Janet D Siliciano, and Robert F Siliciano. Influence of host gene transcription level and orientation on HIV-1 latency in a primary-cell model. *J Virol*, 85(11):5384–5393, Jun 2011. doi: 10.1128/JVI.02536-10. URL http://dx.doi.org/10.1128/JVI.02536-10.

[246] Matthew J Pace, Erin H Graf, Luis M Agosto, Angela M Mexas, Frances Male, Troy Brady, Frederic D Bushman, and Una O'Doherty. Directly infected resting CD4+ T cells can produce HIV Gag without spreading infection in a model of HIV latency. *PLoS Pathog*, 8(7):e1002818, Jul 2012. doi: 10.1371/journal.ppat.1002818. URL http://dx.doi.org/10.1371/journal.ppat.1002818.

[247] Tina Lenasi, Xavier Contreras, and B. Matija Peterlin. Transcriptional in-

terference antagonizes proviral gene expression to promote HIV latency. *Cell Host Microbe*, 4(2):123–133, Aug 2008. doi: 10.1016/j.chom.2008.05.016. URL http://dx.doi.org/10.1016/j.chom.2008.05.016.

[248] Yefei Han, Yijie B Lin, Wenfeng An, Jie Xu, Hung-Chih Yang, Karen O'Connell, Dominic Dordai, Jef D Boeke, Janet D Siliciano, and Robert F Siliciano. Orientation-dependent regulation of integrated HIV-1 expression by host gene transcriptional readthrough. *Cell Host Microbe*, 4(2):134–146, Aug 2008. doi: 10.1016/j.chom.2008.06.008. URL http://dx.doi.org/10.1016/j.chom.2008.06.008.

[249] Liang Shan, Kai Deng, Neeta S Shroff, Christine M Durand, S. Alireza Rabi, Hung-Chih Yang, Hao Zhang, Joseph B Margolick, Joel N Blankson, and Robert F Siliciano. Stimulation of HIV-1-specific cytolytic T lymphocytes facilitates elimination of latent viral reservoir after virus reactivation. *Immunity*, 36(3):491–501, Mar 2012. doi: 10.1016/j.immuni.2012.01.014. URL http://dx.doi.org/10.1016/j.immuni.2012.01.014.

[250] Daniela Boehm, Vincenzo Calvanese, Roy D Dar, Sifei Xing, Sebastian Schroeder, Laura Martins, Katherine Aull, Pao-Chen Li, Vicente Planelles, James E Bradner, Ming-Ming Zhou, Robert F Siliciano, Leor Weinberger, Eric Verdin, and Melanie Ott. BET bromodomain-targeting compounds reactivate HIV from latency via a Tat-independent mechanism. *Cell Cycle*, 12(3):452–462, Feb 2013. doi: 10.4161/cc.23309. URL http://dx.doi.org/10.4161/cc.23309.

[251] Andrea Savarino, Antonello Mai, Sandro Norelli, Sary El Daker, Sergio Valente, Dante Rotili, Lucia Altucci, Anna Teresa Palamara, and Enrico Garaci. "shock and kill" effects of class I-selective histone deacetylase inhibitors in combination with the glutathione synthesis inhibitor buthionine sulfoximine in cell line models for HIV-1 quiescence. *Retrovirology*, 6:52, 2009. doi: 10.1186/1742-4690-6-52. URL http://dx.doi.org/10.1186/1742-4690-6-52.

[252] N. M. Archin, A. L. Liberty, A. D. Kashuba, S. K. Choudhary, J. D. Kuruc, A. M. Crooks, D. C. Parker, E. M. Anderson, M. F. Kearney, M. C. Strain, D. D. Richman, M. G. Hudgens, R. J. Bosch, J. M. Coffin, J. J. Eron, D. J. Hazuda, and D. M. Margolis. Administration of vorinostat disrupts HIV-1 latency in patients on antiretroviral therapy. *Nature*, 487(7408):482–485, Jul 2012. doi: 10.1038/nature11286. URL http://dx.doi.org/10.1038/nature11286.

[253] Alberto Bosque and Vicente Planelles. Induction of HIV-1 latency and reactivation in primary memory CD4+ T cells. *Blood*, 113(1):58–65, Jan 2009. doi: 10.1182/blood-2008-07-168393. URL http://dx.doi.org/10.1182/blood-2008-07-168393.

[254] Alberto Bosque and Vicente Planelles. Studies of HIV-1 latency in an ex vivo model that uses primary central memory T cells. *Methods*, 53(1):54–61, Jan

2011. doi: 10.1016/j.ymeth.2010.10.002. URL http://dx.doi.org/10.1016/j.ymeth.2010.10.002.

[255] Xiaolin Wu, Yuan Li, Bruce Crise, and Shawn M Burgess. Transcription start regions in the human genome are favored targets for MLV integration. *Science*, 300(5626):1749–1751, Jun 2003. doi: 10.1126/science.1083413. URL http://dx.doi.org/10.1126/science.1083413.

[256] Rick S Mitchell, Brett F Beitzel, Astrid R W Schroder, Paul Shinn, Huaming Chen, Charles C Berry, Joseph R Ecker, and Frederic D Bushman. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol*, 2(8):e234, Aug 2004. doi: 10.1371/journal.pbio.0020234. URL http://dx.doi.org/10.1371/journal.pbio.0020234.

[257] Charles Berry, Sridhar Hannenhalli, Jeremy Leipzig, and Frederic D Bushman. Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput Biol*, 2(11):e157, Nov 2006. doi: 10.1371/journal.pcbi.0020157. URL http://dx.doi.org/10.1371/journal.pcbi.0020157.

[258] Gary P Wang, Angela Ciuffi, Jeremy Leipzig, Charles C Berry, and Frederic D Bushman. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res*, 17(8):1186–1194, Aug 2007. doi: 10.1101/gr.6286907. URL http://dx.doi.org/10.1101/gr.6286907.

[259] H. Mochizuki, J. P. Schwartz, K. Tanaka, R. O. Brady, and J. Reiser. High-titer human immunodeficiency virus type 1-based vector systems for gene delivery into nondividing cells. *J Virol*, 72(11):8873–8883, Nov 1998. URL http://jvi.asm.org/content/72/11/8873.abstract.

[260] Yefei Han, Kara Lassen, Daphne Monie, Ahmad R Sedaghat, Shino Shimoji, Xiao Liu, Theodore C Pierson, Joseph B Margolick, Robert F Siliciano, and Janet D Siliciano. Resting CD4+ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. *J Virol*, 78(12):6122–6133, Jun 2004. doi: 10.1128/JVI.78.12.6122-6133.2004. URL http://dx.doi.org/10.1128/JVI.78.12.6122-6133.2004.

[261] Gabriela Plesa, Jihong Dai, Cliff Baytop, James L Riley, Carl H June, and Una O'Doherty. Addition of deoxynucleosides enhances human immunodeficiency virus type 1 integration and 2LTR formation in resting CD4+ T cells. *J Virol*, 81(24):13938–13942, Dec 2007. doi: 10.1128/JVI.01745-07. URL http://dx.doi.org/10.1128/JVI.01745-07.

[262] Nirav Malani. hiReadsProcessor R package. URL http://github.com/malnirav/hiReadsProcessor.

[263] Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, 28(8):817–825, Aug 2010. doi: 10.1038/nbt.1662. URL http://dx.doi.org/10.1038/nbt.1662.

[264] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC genome browser database: update 2006. *Nucleic Acids Res*, 34(Database issue):D590–D598, Jan 2006. doi: 10.1093/nar/gkj144. URL http://dx.doi.org/10.1093/nar/gkj144.

[265] Kate R Rosenbloom, Cricket A Sloan, Venkat S Malladi, Timothy R Dreszer, Katrina Learned, Vanessa M Kirkup, Matthew C Wong, Morgan Maddren, Ruihua Fang, Steven G Heitner, Brian T Lee, Galt P Barber, Rachel A Harte, Mark Diekhans, Jeffrey C Long, Steven P Wilder, Ann S Zweig, Donna Karolchik, Robert M Kuhn, David Haussler, and W. James Kent. ENCODE data in the UCSC genome browser: year 5 update. *Nucleic Acids Res*, 41(D1):D56–D63, Jan 2013. doi: 10.1093/nar/gks1172. URL http://dx.doi.org/10.1093/nar/gks1172.

[266] Jihye Han, Sin-Gi Park, Jae-Bum Bae, JungKyoon Choi, Jae-Myun Lyu, Sung Hee Park, Hei Sung Kim, Young-Joon Kim, Sangsoo Kim, and Tae-Yoon Kim. The characteristics of genome-wide DNA methylation in naïve CD4+ T cells of patients with psoriasis or atopic dermatitis. *Biochem Biophys Res Commun*, 422(1):157–163, May 2012. doi: 10.1016/j.bbrc.2012.04.128. URL http://dx.doi.org/10.1016/j.bbrc.2012.04.128.

[267] Laurence R Meyer, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Robert M Kuhn, Matthew Wong, Cricket A Sloan, Kate R Rosenbloom, Greg Roe, Brooke Rhead, Brian J Raney, Andy Pohl, Venkat S Malladi, Chin H Li, Brian T Lee, Katrina Learned, Vanessa Kirkup, Fan Hsu, Steve Heitner, Rachel A Harte, Maximilian Haeussler, Luvina Guruvadoo, Mary Goldman, Belinda M Giardine, Pauline A Fujita, Timothy R Dreszer, Mark Diekhans, Melissa S Cline, Hiram Clawson, Galt P Barber, David Haussler, and W. James Kent. The UCSC genome browser database: extensions and updates 2013. *Nucleic Acids Res*, 41(D1): D64–D69, Jan 2013. doi: 10.1093/nar/gks1048. URL http://dx.doi.org/10.1093/nar/gks1048.

[268] Zhibin Wang, Chongzhi Zang, Jeffrey A Rosenfeld, Dustin E Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Weiqun Peng, Michael Q Zhang, and Keji Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*, 40(7):897–903, Jul 2008. doi: 10.1038/ng.154. URL http://dx.doi.org/10.1038/ng.154.

[269] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, May 2007. doi: 10.1016/j.cell.2007.05.009. URL http://dx.doi.org/10.1016/j.cell.2007.05.009.

[270] Zhibin Wang, Chongzhi Zang, Kairong Cui, Dustin E Schones, Artem Barski, Weiqun Peng, and Keji Zhao. Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell*, 138(5):1019–1031, Sep 2009. doi: 10.1016/j.cell.2009.06.049. URL http://dx.doi.org/10.1016/j.cell.2009.06.049.

[271] Dustin E Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, Mar 2008. doi: 10.1016/j.cell.2008.02.022. URL http://dx.doi.org/10.1016/j.cell.2008.02.022.

[272] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/.

[273] I. H. Greger, F. Demarchi, M. Giacca, and N. J. Proudfoot. Transcriptional interference perturbs the binding of Sp1 to the HIV-1 promoter. *Nucleic Acids Res*, 26(5):1294–1301, Mar 1998. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC147389/.

[274] Alex De Marco, Chiara Biancotto, Anna Knezevich, Paolo Maiuri, Chiara Vardabasso, and Alessandro Marcello. Intragenic transcriptional cis-activation of the human immunodeficiency virus 1 does not result in allele-specific inhibition of the endogenous gene. *Retrovirology*, 5:98, 2008. doi: 10.1186/1742-4690-5-98. URL http://dx.doi.org/10.1186/1742-4690-5-98.

[275] J. S. Waye and H. F. Willard. Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. *Nucleic Acids Res*, 15(18):7549–7569, Sep 1987. doi: 10.1093/nar/15.18.7549.

[276] E. Verdin, P. Paras, and C. Van Lint. Chromatin disruption in the promoter of human immunodeficiency virus type 1 during transcriptional activation. *EMBO J*, 12(8):3249–3259, Aug 1993. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC413592/.

[277] C. Van Lint, S. Emiliani, M. Ott, and E. Verdin. Transcriptional activation and chromatin remodeling of the HIV-1 promoter in response to histone acetylation. *EMBO J*, 15(5):1112–1120, Mar 1996. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC450009/.

[278] Kara G Lassen, Kasra X Ramyar, Justin R Bailey, Yan Zhou, and Robert F Siliciano. Nuclear retention of multiply spliced HIV-1 RNA in resting CD4+ T cells. *PLoS Pathog*, 2(7):e68, Jul 2006. doi: 10.1371/journal.ppat.0020068. URL http://dx.doi.org/10.1371/journal.ppat.0020068.

[279] Mariacarolina Dieudonné, Paolo Maiuri, Chiara Biancotto, Anna Knezevich, Anna Kula, Marina Lusic, and Alessandro Marcello. Transcriptional competence of the integrated HIV-1 provirus at the nuclear periphery. *EMBO J*, 28(15):2231–2243, Aug 2009. doi: 10.1038/emboj.2009.141. URL http://dx.doi.org/10.1038/emboj.2009.141.

[280] Robert F Siliciano and Warner C Greene. HIV latency. *Cold Spring Harb Perspect Med*, 1(1):a007096, Sep 2011. doi: 10.1101/cshperspect.a007096. URL http://dx.doi.org/10.1101/cshperspect.a007096.

[281] Marina Lusic, Bruna Marini, Hashim Ali, Bojana Lucic, Roberto Luzzati, and Mauro Giacca. Proximity to PML nuclear bodies regulates HIV-1 latency in CD4+ T cells. *Cell Host Microbe*, 13(6):665–677, Jun 2013. doi: 10.1016/j.chom.2013.05.006. URL http://dx.doi.org/10.1016/j.chom.2013.05.006.

[282] L. M. Mansky and H. M. Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol*, 69(8):5087–5094, Aug 1995. URL http://jvi.asm.org/content/69/8/5087.short.

[283] Christopher J L. Murray, Katrina F. Ortblad, Caterina Guinovart, Stephen S. Lim, Timothy M. Wolock, D Allen Roberts, Emily A. Dansereau, Nicholas Graetz, Ryan M. Barber, Jonathan C. Brown, Haidong Wang, Herbert C. Duber, Mohsen Naghavi, Daniel Dicker, Lalit Dandona, Joshua A. Salomon, Kyle R. Heuton, Kyle Foreman, David E. Phillips, Thomas D. Fleming, Abraham D. Flaxman, Bryan K. Phillips, Elizabeth K. Johnson, Megan S. Coggeshall, Foad Abd-Allah, Semaw Ferede Abera, Jerry P. Abraham, Ibrahim Abubakar, Laith J. Abu-Raddad, Niveen Me Abu-Rmeileh, Tom Achoki, Austine Olufemi Adeyemo, Arsène Kouablan Adou, José C. Adsuar, Emilie Elisabet Agardh, Dickens Akena, Mazin J. Al Kahbouri, Deena Alasfoor, Mohammed I. Albittar, Gabriel Alcalá-Cerra, Miguel Angel Alegretti, Zewdie Aderaw Alemu, Rafael Alfonso-Cristancho, Samia Alhabib, Raghib Ali, Francois Alla, Peter J. Allen, Ubai Alsharif, Elena Alvarez, Nelson Alvis-Guzman, Adansi A. Amankwaa, Azmeraw T. Amare, Hassan Amini, Walid Ammar, Benjamin O. Anderson, Carl Abelardo T. Antonio, Palwasha Anwari, Johan Arnlöv, Valentina S Arsic Arsenijevic, Ali Artaman, Rana J. Asghar, Reza Assadi, Lydia S. Atkins, Alaa Badawi, Kalpana Balakrishnan, Amitava Banerjee, Sanjay Basu, Justin Beardsley, Tolesa Bekele, Michelle L. Bell, Eduardo Bernabe, Tariku Jibat Beyene, Neeraj Bhala, Ashish Bhalla, Zulfiqar A. Bhutta, Aref Bin Abdulhak, Agnes Binagwaho, Jed D. Blore, Berrak Bora Basara, Dipan Bose, Michael Brainin, Nicholas Breitborde, Carlos A. Castañeda-Orjuela, Ferrán Catalá-López, Vineet K. Chadha, Jung-Chen Chang, Peggy Pei-Chia Chi-

ang, Ting-Wu Chuang, Mercedes Colomar, Leslie Trumbull Cooper, Cyrus Cooper, Karen J. Courville, Benjamin C. Cowie, Michael H. Criqui, Rakhi Dandona, Anand Dayama, Diego De Leo, Louisa Degenhardt, Borja Del Pozo-Cruz, Kebede Deribe, Don C. Des Jarlais, Muluken Dessalegn, Samath D. Dharmaratne, Uur Dilmen, Eric L. Ding, Tim R. Driscoll, Adnan M. Durrani, Richard G. Ellenbogen, Sergey Petrovich Ermakov, Alireza Esteghamati, Emerito Jose A. Faraon, Farshad Farzadfar, Seyed-Mohammad Fereshtehnejad, Daniel Obadare Fijabi, Mohammad H. Forouzanfar, Urbano Fra Paleo, Lynne Gaffikin, Amiran Gamkrelidze, Fortuné Gbètoho Gankpé, Johanna M. Geleijnse, Bradford D. Gessner, Katherine B. Gibney, Ibrahim Abdelmageem Mohamed Ginawi, Elizabeth L. Glaser, Philimon Gona, Atsushi Goto, Hebe N. Gouda, Harish Chander Gugnani, Rajeev Gupta, Rahul Gupta, Nima Hafezi-Nejad, Randah Ribhi Hamadeh, Mouhanad Hammami, Graeme J. Hankey, Hilda L. Harb, Josep Maria Haro, Rasmus Havmoeller, Simon I. Hay, Mohammad T. Hedayati, Ileana B Heredia Pi, Hans W. Hoek, John C. Hornberger, H Dean Hosgood, Peter J. Hotez, Damian G. Hoy, John J. Huang, Kim M. Iburg, Bulat T. Idrisov, Kaire Innos, Kathryn H. Jacobsen, Panniyammakal Jeemon, Paul N. Jensen, Vivekanand Jha, Guohong Jiang, Jost B. Jonas, Knud Juel, Haidong Kan, Ida Kankindi, Nadim E. Karam, André Karch, Corine Kakizi Karema, Anil Kaul, Norito Kawakami, Dhruv S. Kazi, Andrew H. Kemp, Andre Pascal Kengne, Andre Keren, Maia Kereselidze, Yousef Saleh Khader, Shams Eldin Ali Hassan Khalifa, Ejaz Ahmed Khan, Young-Ho Khang, Irma Khonelidze, Yohannes Kinfu, Jonas M. Kinge, Luke Knibbs, Yoshihiro Kokubo, S. Kosen, Barthelemy Kuate Defo, Veena S. Kulkarni, Chanda Kulkarni, Kaushalendra Kumar, Ravi B. Kumar, G Anil Kumar, Gene F. Kwan, Taavi Lai, Arjun Lakshmana Balaji, Hilton Lam, Qing Lan, Van C. Lansingh, Heidi J. Larson, Anders Larsson, Jong-Tae Lee, James Leigh, Mall Leinsalu, Ricky Leung, Yichong Li, Yongmei Li, Graça Maria Ferreira De Lima, Hsien-Ho Lin, Steven E. Lipshultz, Shiwei Liu, Yang Liu, Belinda K. Lloyd, Paulo A. Lotufo, Vasco Manuel Pedro Machado, Jennifer H. Maclachlan, Carlos Magis-Rodriguez, Marek Majdan, Christopher Chabila Mapoma, Wagner Marcenes, Melvin Barrientos Marzan, Joseph R. Masci, Mohammad Taufiq Mashal, Amanda J. Mason-Jones, Bongani M. Mayosi, Tasara T. Mazorodze, Abigail Cecilia Mckay, Peter A. Meaney, Man Mohan Mehndiratta, Fabiola Mejia-Rodriguez, Yohannes Adama Melaku, Ziad A. Memish, Walter Mendoza, Ted R. Miller, Edward J. Mills, Karzan Abdulmuhsin Mohammad, Ali H. Mokdad, Glen Liddell Mola, Lorenzo Monasta, Marcella Montico, Ami R. Moore, Rintaro Mori, Wilkister Nyaora Moturi, Mitsuru Mukaigawara, Kinnari S. Murthy, Aliya Naheed, Kovin S. Naidoo, Luigi Naldi, Vinay Nangia, K M Venkat Narayan, Denis Nash, Chakib Nejjari, Robert G. Nelson, Sudan Prasad Neupane, Charles R. Newton, Marie Ng, Muhammad Imran Nisar, Sandra Nolte, Ole F. Norheim, Vincent Nowaseb, Luke Nyakarahuka, In-Hwan Oh, Takayoshi Ohkubo, Bolajoko O. Olusanya, Saad B. Omer, John Nelson Opio, Orish Ebere Orisakwe, Jeyaraj D. Pandian, Christina Papachristou, Angel J Paternina Caicedo, Scott B. Patten, Vinod K. Paul, Boris Igor Pavlin, Neil Pearce, David M. Pereira, Aslam Pervaiz, Konrad Pesudovs, Max Petzold, Farshad Pourmalek, Dima Qato, Amado D. Quezada, D Alex Quistberg, Anwar Rafay, Kazem

Rahimi, Vafa Rahimi-Movaghar, Sajjad Ur Rahman, Murugesan Raju, Saleem M. Rana, Homie Razavi, Robert Quentin Reilly, Giuseppe Remuzzi, Jan Hendrik Richardus, Luca Ronfani, Nobhojit Roy, Nsanzimana Sabin, Mohammad Yahya Saeedi, Mohammad Ali Sahraian, Genesis May J. Samonte, Monika Sawhney, Ione J C. Schneider, David C. Schwebel, Soraya Seedat, Sadaf G. Sepanlou, Edson E. Servan-Mori, Sara Sheikhbahaei, Kenji Shibuya, Hwashin Hyun Shin, Ivy Shiue, Rupak Shivakoti, Inga Dora Sigfusdottir, Donald H. Silberberg, Andrea P. Silva, Edgar P. Simard, Jasvinder A. Singh, Vegard Skirbekk, Karen Sliwa, Samir Soneji, Sergey S. Soshnikov, Chandrashekhar T. Sreeramareddy, Vasiliki Kalliopi Stathopoulou, Konstantinos Stroumpoulis, Soumya Swaminathan, Bryan L. Sykes, Karen M. Tabb, Roberto Tchio Talongwa, Eric Yeboah Tenkorang, Abdullah Sulieman Terkawi, Alan J. Thomson, Andrew L. Thorne-Lyman, Jeffrey A. Towbin, Jefferson Traebert, Bach X. Tran, Zacharie Tsala Dimbuene, Miltiadis Tsilimbaris, Uche S. Uchendu, Kingsley N. Ukwaja, Selen Begüm Uzun, Andrew J. Vallely, Tommi J. Vasankari, N. Venketasubramanian, Francesco S. Violante, Vasiliy Victorovich Vlassov, Stein Emil Vollset, Stephen Waller, Mitchell T. Wallin, Linhong Wang, XiaoRong Wang, Yanping Wang, Scott Weichenthal, Elisabete Weiderpass, Robert G. Weintraub, Ronny Westerman, Richard A. White, James D. Wilkinson, Thomas Neil Williams, Solomon Meseret Woldeyohannes, John Q. Wong, Gelin Xu, Yang C. Yang, Yuichiro Yano, Gokalp Kadri Yentur, Paul Yip, Naohiro Yonemoto, Seok-Jun Yoon, Mustafa Younis, Chuanhua Yu, Kim Yun Jin, Maysaa El Sayed Zaki, Yong Zhao, Yingfeng Zheng, Maigeng Zhou, Jun Zhu, Xiao Nong Zou, Alan D. Lopez, and Theo Vos. Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*, 384(9947):1005–1070, Sep 2014. doi: 10.1016/S0140-6736(14)60844-8. URL http://dx.doi.org/10.1016/S0140-6736(14)60844-8.

[284] Kimberly A. Sollis, Pieter W. Smit, Susan Fiscus, Nathan Ford, Marco Vitoria, Shaffiq Essajee, David Barnett, Ben Cheng, Suzanne M. Crowe, Thomas Denny, Alan Landay, Wendy Stevens, Vincent Habiyambere, Jos Perrins, and Rosanna W. Peeling. Systematic review of the performance of HIV viral load technologies on plasma samples. *PLoS One*, 9(2):e85869, 2014. doi: 10.1371/journal.pone.0085869. URL http://dx.doi.org/10.1371/journal.pone.0085869.

[285] Changchun Liu, Michael Mauk, Robert Gross, Frederic D. Bushman, Paul H. Edelstein, Ronald G. Collman, and Haim H. Bau. Membrane-based, sedimentation-assisted plasma separator for point-of-care applications. *Anal Chem*, 85(21): 10463–10470, Nov 2013. doi: 10.1021/ac402459h. URL http://dx.doi.org/10.1021/ac402459h.

[286] Kelly A. Curtis, Donna L. Rudolph, Irene Nejad, Jered Singleton, Andy Beddoe, Bernhard Weigl, Paul LaBarre, and S Michele Owen. Isothermal amplification using a chemical heating device for point-of-care detection of HIV-1. *PLoS One*, 7(2):e31432, 2012. doi: 10.1371/journal.pone.0031432. URL http://dx.doi.org/10.1371/journal.pone.0031432.

[287] T. Notomi, H. Okayama, H. Masubuchi, T. Yonekawa, K. Watanabe, N. Amino, and T. Hase. Loop-mediated isothermal amplification of DNA. *Nucleic Acids Res*, 28 (12):E63, Jun 2000. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102748/.

[288] Kelly A. Curtis, Donna L. Rudolph, and S Michele Owen. Rapid detection of HIV-1 by reverse-transcription, loop-mediated isothermal amplification (RT-LAMP). *J Virol Methods*, 151(2):264–270, Aug 2008. doi: 10.1016/j.jviromet.2008.04.011. URL http://dx.doi.org/10.1016/j.jviromet.2008.04.011.

[289] Kelly A. Curtis, Donna L. Rudolph, and S Michele Owen. Sequence-specific detection method for reverse transcription, loop-mediated isothermal amplification of HIV-1. *J Med Virol*, 81(6):966–972, Jun 2009. doi: 10.1002/jmv.21490. URL http://dx.doi.org/10.1002/jmv.21490.

[290] Yalan Zeng, Xiaoguang Zhang, Kai Nie, Xiong Ding, Brian Z. Ring, Lanying Xu, Lei Dai, Xiying Li, Wei Ren, Lei Shi, and Xuejun Ma. Rapid quantitative detection of human immunodeficiency virus type 1 by a reverse transcription-loop-mediated isothermal amplification assay. *Gene*, 541(2):123–128, May 2014. doi: 10.1016/j.gene.2014.03.015. URL http://dx.doi.org/10.1016/j.gene.2014.03.015.

[291] Norimitsu Hosaka, Nicaise Ndembi, Azumi Ishizaki, Seiji Kageyama, Kei Numazaki, and Hiroshi Ichimura. Rapid detection of human immunodeficiency virus type 1 group M by a reverse transcription-loop-mediated isothermal amplification assay. *J Virol Methods*, 157(2):195–199, May 2009. doi: 10.1016/j.jviromet.2009.01.004. URL http://dx.doi.org/10.1016/j.jviromet.2009.01.004.

[292] Kelly A. Curtis, Philip L. Niedzwiedz, Ae S. Youngpairoj, Donna L. Rudolph, and S Michele Owen. Real-time detection of HIV-2 by reverse transcription-loop-mediated isothermal amplification. *J Clin Microbiol*, 52(7):2674–2676, Jul 2014. doi: 10.1128/JCM.00935-14. URL http://dx.doi.org/10.1128/JCM.00935-14.

[293] Carla Kuiken, Hyejin Yoon, Werner Abfalterer, Brian Gaschen, Chienchi Lo, and Bette Korber. Viral genome analysis and knowledge management. *Methods Mol Biol*, 939:253–261, 2013. doi: 10.1007/978-1-62703-107-3_16. URL http://dx.doi.org/10.1007/978-1-62703-107-3_16.

[294] Mark Manak, Silvana Sina, Bharathi Anekella, Indira Hewlett, Eric Sanders-Buell, Viswanath Ragupathy, Jerome Kim, Marion Vermeulen, Susan L. Stramer, Ester Sabino, Piotr Grabarczyk, Nelson Michael, Sheila Peel, Patricia Garrett, Sodsai Tovanabutra, Michael P. Busch, and Marco Schito. Pilot studies for development of an HIV subtype panel for surveillance of global diversity. *AIDS Res Hum Retroviruses*, 28(6):594–606, Jun 2012. doi: 10.1089/AID.2011.0271. URL http://dx.doi.org/10.1089/AID.2011.0271.

[295] J. Louwagie, F. E. McCutchan, M. Peeters, T. P. Brennan, E. Sanders-Buell, G. A. Eddy, G. van der Groen, K. Fransen, G. M. Gershy-Damet, and R. Deleys. Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. *AIDS*, 7(6):769–780, Jun 1993. URL http://www.ncbi.nlm.nih.gov/pubmed/8363755.

[296] L. Buonaguro, M. L. Tornesello, and F. M. Buonaguro. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. *J Virol*, 81(19):10209–10219, Oct 2007. doi: 10.1128/JVI.00872-07. URL http://dx.doi.org/10.1128/JVI.00872-07.

[297] Nicholas F. Parrish, Feng Gao, Hui Li, Elena E. Giorgi, Hannah J. Barbian, Erica H. Parrish, Lara Zajic, Shilpa S. Iyer, Julie M. Decker, Amit Kumar, Bhavna Hora, Anna Berg, Fangping Cai, Jennifer Hopper, Thomas N. Denny, Haitao Ding, Christina Ochsenbauer, John C. Kappes, Rachel P. Galimidi, Anthony P West, Jr, Pamela J. Bjorkman, Craig B. Wilen, Robert W. Doms, Meagan O'Brien, Nina Bhardwaj, Persephone Borrow, Barton F. Haynes, Mark Muldoon, James P. Theiler, Bette Korber, George M. Shaw, and Beatrice H. Hahn. Phenotypic properties of transmitted founder HIV-1. *Proc Natl Acad Sci U S A*, 110(17):6626–6633, Apr 2013. doi: 10.1073/pnas.1304288110. URL http://dx.doi.org/10.1073/pnas.1304288110.

[298] A. G. Abimiku, T. L. Stern, A. Zwandor, P. D. Markham, C. Calef, S. Kyari, W. C. Saxinger, R. C. Gallo, M. Robert-Guroff, and M. S. Reitz. Subgroup G HIV type 1 isolates from Nigeria. *AIDS Res Hum Retroviruses*, 10(11):1581–1583, Nov 1994. URL http://www.ncbi.nlm.nih.gov/pubmed/7888214.

[299] Zhang, Chung, and Oldenburg. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J Biomol Screen*, 4(2):67–73, 1999. doi: 10.1177/108705719900400206.

[300] Changchun Liu, Eran Geva, Michael Mauk, Xianbo Qiu, William R. Abrams, Daniel Malamud, Kelly Curtis, S Michele Owen, and Haim H. Bau. An isothermal amplification reactor with an integrated isolation membrane for point-of-care detection of infectious diseases. *Analyst*, 136(10):2069–2076, May 2011. doi: 10.1039/c1an00007a. URL http://dx.doi.org/10.1039/c1an00007a.

[301] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40:1413–1415, 2008. doi: 10.1038/ng.259. URL http://www.nature.com/ng/journal/v40/n12/full/ng.259.html.

[302] Franco Pagani, Michela Raponi, and Francisco E. Baralle. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc Natl Acad Sci U S A*, 102(18):6368–6372, May 2005. doi: 10.1073/pnas.0502288102. URL http://dx.doi.org/10.1073/pnas.0502288102.

[303] Kai Wang, Rasmus Wernersson, and Søren Brunak. The strength of intron donor splice sites in human genes displays a bell-shaped pattern. *Bioinformatics*, 27(22):3079–3084, Nov 2011. doi: 10.1093/bioinformatics/btr532. URL http://dx.doi.org/10.1093/bioinformatics/btr532.

[304] Martin Lützelberger, Line S Reinert, Atze T Das, Ben Berkhout, and Jørgen Kjems. A novel splice donor site in the gag-pol gene is required for HIV-1 RNA stability. *J Biol Chem*, 281(27):18644–18651, Jul 2006. doi: 10.1074/jbc.M513698200. URL http://dx.doi.org/10.1074/jbc.M513698200.

[305] J. Salfeld, H. G. Gttlinger, R. A. Sia, R. E. Park, J. G. Sodroski, and W. A. Haseltine. A tripartite HIV-1 tat-env-rev fusion protein. *EMBO J*, 9(3):965–970, Mar 1990. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC551759/?tool=pubmed.

[306] Joseph A. Jablonski and Massimo Caputi. Role of cellular RNA processing factors in human immunodeficiency virus type 1 mRNA metabolism, replication, and infectivity. *J Virol*, 83(2):981–992, Jan 2009. doi: 10.1128/JVI.01801-08. URL http://dx.doi.org/10.1128/JVI.01801-08.

[307] Anna Tranell, Susanne Tingsborg, Eva Maria Feny, and Stefan Schwartz. Inhibition of splicing by serine-arginine rich protein 55 (SRp55) causes the appearance of partially spliced HIV-1 mRNAs in the cytoplasm. *Virus Res*, 157(1):82–91, Apr 2011. doi: 10.1016/j.virusres.2011.02.010. URL http://dx.doi.org/10.1016/j.virusres.2011.02.010.

[308] Honglin Zhou, Min Xu, Qian Huang, Adam T. Gates, Xiaohua D. Zhang, John C. Castle, Erica Stec, Marc Ferrer, Berta Strulovici, Daria J. Hazuda, and Amy S. Espeseth. Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe*, 4(5):495–504, Nov 2008. doi: 10.1016/j.chom.2008.10.004. URL http://dx.doi.org/10.1016/j.chom.2008.10.004.

[309] Yiping Zhu, Guifang Chen, Fengxiang Lv, Xinlu Wang, Xin Ji, Yihui Xu, Jing Sun, Li Wu, Yong-Tang Zheng, and Guangxia Gao. Zinc-finger antiviral protein inhibits HIV-1 infection by selectively targeting multiply spliced viral mRNAs for degradation. *Proc Natl Acad Sci U S A*, 108(38):15834–15839, Sep 2011. doi: 10.1073/pnas.1101676108. URL http://dx.doi.org/10.1073/pnas.1101676108.

[310] M. J. Saltarelli, E. Hadziyannis, C. E. Hart, J. V. Harrison, B. K. Felber, T. J. Spira, and G. N. Pavlakis. Analysis of human immunodeficiency virus type 1 mRNA splicing patterns during disease progression in peripheral blood mononuclear cells from infected individuals. *AIDS Res Hum Retroviruses*, 12(15):1443–1456, Oct 1996. doi: 10.1089/aid.1996.12.1443.

[311] Elena Delgado, Cristina Carrera, Paloma Nebreda, Aurora Fernndez-Garca, Milagros Pinilla, Valentina Garca, Luca Prez-lvarez, and Michael M Thomson. Identification of new splice sites used for generation of rev transcripts in human

immunodeficiency virus type 1 subtype C primary isolates. *PLoS One*, 7(2): e30574, 2012. doi: 10.1371/journal.pone.0030574. URL http://dx.doi.org/10.1371/journal.pone.0030574.

[312] Paula Grabowski. Alternative splicing takes shape during neuronal development. *Curr Opin Genet Dev*, 21(4):388–394, Aug 2011. doi: 10.1016/j.gde.2011.03.005. URL http://dx.doi.org/10.1016/j.gde.2011.03.005.

[313] Miriam Llorian and Christopher W J. Smith. Decoding muscle alternative splicing. *Curr Opin Genet Dev*, 21(4):380–387, Aug 2011. doi: 10.1016/j.gde.2011.03.006. URL http://dx.doi.org/10.1016/j.gde.2011.03.006.

[314] Joanna Y Ip, Alan Tong, Qun Pan, Justin D Topp, Benjamin J Blencowe, and Kristen W Lynch. Global analysis of alternative splicing during T-cell activation. *RNA*, 13(4):563–572, Apr 2007. doi: 10.1261/rna.457207. URL http://dx.doi.org/10.1261/rna.457207.

[315] Justin D. Topp, Jason Jackson, Alexis A. Melton, and Kristen W. Lynch. A cell-based screen for splicing regulators identifies hnRNP LL as a distinct signal-induced repressor of CD45 variable exon 4. *RNA*, 14(10):2038–2049, Oct 2008. doi: 10.1261/rna.1212008. URL http://dx.doi.org/10.1261/rna.1212008.

[316] Secondo Sonza, Helen P. Mutimer, Kate O'Brien, Philip Ellery, Jane L. Howard, Jonathan H. Axelrod, Nicholas J. Deacon, Suzanne M. Crowe, and Damian F J. Purcell. Selectively reduced tat mRNA heralds the decline in productive human immunodeficiency virus type 1 infection in monocyte-derived macrophages. *J Virol*, 76(24):12611–12621, Dec 2002. URL http://jvi.asm.org/content/76/24/12611.abstract?ijkey=0acf427bc0e3b3f9b3d92663d3cd9cbcd0d9f533&keytype2=tf_ipsecsha.

[317] H. Deng, R. Liu, W. Ellmeier, S. Choe, D. Unutmaz, M. Burkhart, P. Di Marzio, S. Marmon, R. E. Sutton, C. M. Hill, C. B. Davis, S. C. Peiper, T. J. Schall, D. R. Littman, and N. R. Landau. Identification of a major co-receptor for primary isolates of HIV-1. *Nature*, 381(6584):661–666, Jun 1996. doi: 10.1038/381661a0. URL http://dx.doi.org/10.1038/381661a0.

[318] N. R. Landau and D. R. Littman. Packaging system for rapid production of murine leukemia virus vectors with variable tropism. *J Virol*, 66(8):5110–5113, Aug 1992. URL http://jvi.asm.org/content/66/8/5110.abstract?ijkey=231a454ff2c64861cf0114b60c0f2bac80138b1d&keytype2=tf_ipsecsha.

[319] Xiping Wei, Julie M. Decker, Shuyi Wang, Huxiong Hui, John C. Kappes, Xiaoyun Wu, Jesus F. Salazar-Gonzalez, Maria G. Salazar, J Michael Kilby, Michael S. Saag, Natalia L. Komarova, Martin A. Nowak, Beatrice H. Hahn, Peter D. Kwong, and George M. Shaw. Antibody neutralization and escape by HIV-1. *Nature*, 422

(6929):307–312, Mar 2003. doi: 10.1038/nature01470. URL http://dx.doi.org/10.1038/nature01470.

[320] N. Srinivasakumar, N. Chazal, C. Helga-Maria, S. Prasad, M. L. Hammarskjöld, and D. Rekosh. The effect of viral regulatory protein expression on gene delivery by human immunodeficiency virus type 1 vectors produced in stable packaging cell lines. *J Virol*, 71(8):5841–5848, Aug 1997. URL http://jvi.asm.org/content/71/8/5841.abstract?ijkey=3558f7baecbb670dbb7ea15e6d57db59f537fbcd&keytype2=tf_ipsecsha.

[321] D. C. Shugars, M. S. Smith, D. H. Glueck, P. V. Nantermet, F. Seillier-Moiseiwitsch, and R. Swanstrom. Analysis of human immunodeficiency virus type 1 nef gene sequences present in vivo. *J Virol*, 67(8):4639–4650, Aug 1993. URL http://jvi.asm.org/content/67/8/4639.abstract?ijkey=f8c681831993b7eef45161cdbfa2d4936d66ae52&keytype2=tf_ipsecsha.

[322] Kevin J Travers, Chen-Shan Chin, David R Rank, John S Eid, and Stephen W Turner. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res*, 38(15):e159, Aug 2010. doi: 10.1093/nar/gkq543. URL http://dx.doi.org/10.1093/nar/gkq543.

[323] T. A. Thanaraj and F. Clark. Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res*, 29(12):2581–2593, Jun 2001. doi: 10.1093/nar/29.12.2581. URL http://dx.doi.org/10.1093/nar/29.12.2581.

[324] M. Aebi, H. Hornig, R. A. Padgett, J. Reiser, and C. Weissmann. Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. *Cell*, 47(4):555–565, Nov 1986. URL http://www.sciencedirect.com/science/article/pii/0092867486906203.

[325] M. Burset, I. A. Seledtsov, and V. V. Solovyev. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res*, 28(21):4364–4375, Nov 2000. doi: 10.1093/nar/28.21.4364.

[326] M. Burset, I. A. Seledtsov, and V. V. Solovyev. SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res*, 29(1):255–259, Jan 2001. doi: 10.1093/nar/29.1.255.

[327] Nihar Sheth, Xavier Roca, Michelle L. Hastings, Ted Roeder, Adrian R. Krainer, and Ravi Sachidanandam. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res*, 34(14):3955–3967, 2006. doi: 10.1093/nar/gkl556. URL http://dx.doi.org/10.1093/nar/gkl556.

[328] J. C. Guatelli, T. R. Gingeras, and D. D. Richman. Alternative splice acceptor uti-

lization during human immunodeficiency virus type 1 infection of cultured cells. *J Virol*, 64(9):4093–4098, Sep 1990. URL http://jvi.asm.org/content/64/9/4093.abstract?ijkey=a64601a3b7b53cdb2d4e03d980686acb3ca4c5e5&keytype2=tf_ipsecsha.

[329] C Kuiken, B Foley, T Leitner, C Apetrei, B Hahn, I Mizrachi, J Mullins, A Rambaut, S Wolinsky, and B Korber. HIV sequence compendium 2010. Technical Report LA-UR 10-03684, Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, New Mexio, 2010. URL http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/2010compendium.html.

[330] Christopher Burge, Thomas Tuschl, and Phillip Sharp. Splicing of precursors to mRNAs by the spliceosomes. *Cold Spring Harbor Monograph Archive*, 37, 1999. doi: 10.1101/087969589.37.525. URL http://cshmonographs.org/csh/index.php/monographs/article/view/5123/4220.

[331] Truus E M Abbink and Ben Berkhout. RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor. *J Virol*, 82(6):3090–3098, Mar 2008. doi: 10.1128/JVI.01479-07. URL http://dx.doi.org/10.1128/JVI.01479-07.

[332] K. Verhoef, P. S. Bilodeau, J. L. van Wamel, J. Kjems, C. M. Stoltzfus, and B. Berkhout. Repair of a Rev-minus human immunodeficiency virus type 1 mutant by activation of a cryptic splice site. *J Virol*, 75(7):3495–3500, Apr 2001. doi: 10.1128/JVI.75.7.3495-3500.2001. URL http://dx.doi.org/10.1128/JVI.75.7.3495-3500.2001.

[333] Alan M. Zahler, Christian K. Damgaard, Jorgen Kjems, and Massimo Caputi. SC35 and heterogeneous nuclear ribonucleoprotein A/B proteins bind to a juxtaposed exonic splicing enhancer/exonic splicing silencer element to regulate HIV-1 tat exon 2 splicing. *J Biol Chem*, 279(11):10077–10084, Mar 2004. doi: 10.1074/jbc.M312743200. URL http://dx.doi.org/10.1074/jbc.M312743200.

[334] Brandon F Keele, Elena E Giorgi, Jesus F Salazar-Gonzalez, Julie M Decker, Kimmy T Pham, Maria G Salazar, Chuanxi Sun, Truman Grayson, Shuyi Wang, Hui Li, Xiping Wei, Chunlai Jiang, Jennifer L Kirchherr, Feng Gao, Jeffery A Anderson, Li-Hua Ping, Ronald Swanstrom, Georgia D Tomaras, William A Blattner, Paul A Goepfert, J. Michael Kilby, Michael S Saag, Eric L Delwart, Michael P Busch, Myron S Cohen, David C Montefiori, Barton F Haynes, Brian Gaschen, Gayathri S Athreya, Ha Y Lee, Natasha Wood, Cathal Seoighe, Alan S Perelson, Tanmoy Bhattacharya, Bette T Korber, Beatrice H Hahn, and George M Shaw. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci USA*, 105(21):7552–7557, May 2008. doi: 10.1073/pnas.0802203105. URL http://dx.doi.org/10.1073/pnas.0802203105.

[335] Jesus F Salazar-Gonzalez, Maria G Salazar, Brandon F Keele, Gerald H Learn,

Elena E Giorgi, Hui Li, Julie M Decker, Shuyi Wang, Joshua Baalwa, Matthias H Kraus, Nicholas F Parrish, Katharina S Shaw, M. Brad Guffey, Katharine J Bar, Katie L Davis, Christina Ochsenbauer-Jambor, John C Kappes, Michael S Saag, Myron S Cohen, Joseph Mulenga, Cynthia A Derdeyn, Susan Allen, Eric Hunter, Martin Markowitz, Peter Hraber, Alan S Perelson, Tanmoy Bhattacharya, Barton F Haynes, Bette T Korber, Beatrice H Hahn, and George M Shaw. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J Exp Med*, 206(6):1273–1289, Jun 2009. doi: 10.1084/jem.20090378. URL http://dx.doi.org/10.1084/jem.20090378.