# Investigation of HERV-K (HML-2) expression during HIV-1 infection

A thesis submitted by

Neeru Bhardwaj

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in

Molecular Microbiology

MERGE-ID

TUFTS UNIVERSITY

Sackler School of Graduate Biomedical Sciences

August, 2015

Advisor: John Coffin

Thesis Chair: Naomi Rosenberg

**Abstract**

Human endogenous retrovirus group K (HERV-K) proviruses are among the limited number of human endogenous retroviral elements to retain coding sequence. Of interest, expression from the Human Mouse Mammary Tumor Virus-like 2 (HML-2) subgroup of HERV-K proviruses has been widely associated with disease, including different types of cancers as well as in HIV-1 infection. In particular, recent studies have suggested that HML-2 expression may play a role in the pathogenesis of HIV-1 infection. RNA from the HML-2 subgroup was reported to be highly expressed at the cellular level and detectable in the plasma of HIV-1 infected patients, suggestive of virion production and potentially replication.

In order to investigate this phenomenon, an HML-2 specific quantitative PCR assay was developed, which detects 51 of the >90 known HML-2 proviruses in the human genome. Plasma and peripheral blood mononuclear cells (PBMCs) from HIV-negative controls and HIV-1 infected patients were collected for analysis of HML-2 RNA expression. Contrary to previous reports, high levels of HML-2 RNA were not detected in the plasma of HIV-1 infected patients, but a significant increase of HML-2 RNA in total PBMCs was observed. The level of HML-2 expression in PBMCs did not appear to be related to patient use of antiretrovirals or to HIV-1 plasma RNA, cellular RNA or cellular DNA levels. To investigate the source of expressed HML-2 RNA, patient PBMCs were sorted into CD3+CD4+ T cells, CD3+CD8+ T cells, CD3-CD14+ monocytes and CD3-CD20+ B cell subsets and then analyzed for HML-2 RNA levels using the quantitative PCR assay. No single cell subset was enriched for HML-2 RNA

expression in HIV-1 infected patients, but there was substantial variability in the level of HML-2 expression dependent on the cell type.

In order to understand the potential impact of HML-2 expression, an RNAseq methodology was developed to annotate expression at the proviral level. Prior to use on the HIV-1 infected and uninfected populations, this RNASeq strategy was implemented using the teratocarcinoma cell line Tera-1, known to express high levels of cellular HML-2 RNA and produce non-infectious HML-2 virions. RNASeq effectively discerned the proviral expression pattern of this cell line and was capable of identifying the core expressed transcripts, which originated from two proviruses located on chromosome 22 (chr 22q11.21 and chr 22q11.23). Interestingly, only one of these proviral transcripts appeared to be packaged into virions at a high level, suggesting a preference for recently integrated proviruses to be packaged into virions over those from other highly expressed but older elements.

The validated RNASeq approach was used to investigate the HML-2 profile of PBMCs collected from HIV-1 infected and uninfected subjects. Bioinformatic analysis revealed that HML-2 expression profiles between these populations are remarkably similar in composition, though not in magnitude of expression. Increased overall HML-2 expression detected in the HIV-1 population appears to be driven by the higher expression of the provirus on chr 1q22. However, as this provirus does not maintain open reading frames for retroviral genes *gag*, *pro*, *pol* and *env*, it is unlikely that it leads to production of retroviral particles or participates in productive recombination events that could lead to replicating HML-2 virus. Thus, based on these studies, HML-2 expression during HIV-1 infection is not predicted to have a direct pathogenic effect.

**Acknowledgments**

I expect that I will not be writing another book soon, so I would like to dedicate my thesis to my parents, Bhu Dev and Poonam Bhardwaj. They sacrificed endlessly to ensure that I had the best education available and access to opportunities they did not have in their own lives. I am only now starting to grasp all they provided for me in my childhood, and how they continue to support me even now. I would be a much different person if I did not grow up with their examples of dedication, independence and selfless love, and importantly, their unwavering commitment to delicious food.

I feel very lucky to have been a part of the Coffin lab in my graduate career. In John, I found an excellent advisor and a model for how to be a diligent scientist, but also a leader. He gave me the opportunity to research, hypothesize freely, attend conferences and meet and work with the best scientists in the field. He always made time to chat and is right about almost everything, which was invaluable to me and provided great direction.

In John's lab, I found a motley crew of fellow students. The Coffin lab is full of effusively warm, hilarious and supportive personalities, all of whom made every day a joy for me, illustrated by my cackling down the halls of Jaharis 4. Zach Williams, my bay mate, is intelligent and generous, and my experience at Tufts would not have been as enjoyable without him sitting next to me, talking to himself. He and my other lab mates – Mike Freeman, Joe Holloway, Meagan Montesion, Farrah Roy and John Yoon – make any other work place dull in comparison. I, without a doubt in my mind, worked in an ideal environment, save for the gurgling vacuum pump spewing oil.

Outside of the lab, I had a network of peers, co-workers and collaborators who made my experience at Tufts engaging and successful. In particular, I had wonderful classmates who became friends and would like to recognize Seble Asrat, who started out as my MERGE-ID counterpart but quickly became a sister to me. She is a woman who I willingly worked with for hours in silent library rooms to complete my thesis and qualifier, and with whom I've shared countless coffees, lunches and laughs.

I could not have imagined what I was going to learn when I entered this program. I ended up learning microbiological and analytical skills, of course, but also (some) patience, how to deal with failure and a general acceptance that most things are out of my control. I assume my husband Randy Wurster, who I married during my first year in the program, appreciates the latter half more. Randy is an unofficial member of the Coffin lab and will never forget that 8% of the human genome is made up of endogenous retroviruses. He is my favorite person in the world not just because he is a great cook, which is undoubtedly the only reason why I don't eat bread and cheese every day, but because he is loving, thoughtful, understanding and brilliant in many ways, helping me innumerable times during the course of my PhD.

**Table of Contents**

**List of Tables**

**List of Figures**

**List of Abbreviations**

ACTB: actin beta

ADCC: antibody dependent cell-mediated cytotoxcity

AIDS: acquired immunodeficiency syndrome

AIM2: absent in melanoma 2

ALV: avian leukosis virus

APOBEC: apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like

ART: anti-retroviral therapy

ASRGL1: asparaginase L 1

ATL: adult T-cell leukemia/lymphoma

BSA: bovine serum albumin

CA: capsid

CD: cluster of differentiation

CEBPB: CCAAT/enhancer-binding protein beta

ChIP-Seq: chromosomal immunoprecipitation sequencing

Chr: chromosome

CSF1R: colony stimulating factor 1 receptor

CT-RCC-1: cytotoxic T cell renal cell carcinoma 1

DDX60: DEAD box protein 60

DMSO: dimethyl sulfoxide

DNMT: DNA methyltransferase

DTT: dithiothreitol

EBV: Epstein-Barr virus

EDTA: Ethylenediaminetetraacetic acid

ENCODE: enclycopedia of DNA elements

Env: envelope

ERV: endogenous retrovirus

ESET: ERG-associated protein with SET domain

FBS: fetal bovine serum

FPKM: fragments per kilobase per million mapped reads

GAPDH: glyceraldehyde 3-phosphate dehydrogenase

H3K9: histone 3 lysine (K) 9

H3K27: histone 3 lysine (K) 27

HAART: highly active antiretroviral therapy

HCV: hepatitis C virus

HERV: human endogenous retrovirus

HERV-K: human endogenous retrovirus group K

HHV-8: human herpesvirus 8

HIV: human immunodeficiency virus

HLA: human leukocyte antigen

HML-2: human mouse mammary tumor virus-like group 2

hnRNP: heterogeneous nuclear ribonucleoproteins

HPRT1: hypoxanthine-guanine phosphoribosyltransferase 1

HTDV: human teratocarcinoma derived virus

HTLV: human T-lymphotropic virus

IFI16: interferon gamma inducible protein 16

Il: interleukin

IN: integrase

Inr: initator element

IRB: Institutional Regulatory Board

iSCA: integrase single copy assay

ISD: immunosuppressive domain

IVT: *in vitro* transcription

JSRV: Jaagsiekte sheep retrovirus

KAP1: KRAB associated protein 1

KRAB-ZNF: Krüppel-associated box zinc finger protein

lincRNA: long intergenic noncoding RNA

lincRoR: lincRNA regulator of reprogramming

LNX: ligand of Numb protein X

LOCK: large organized chromatin K9-modification

LOD: limit of detection

LTR: long terminal repeat

MA: matrix

MAPQ: mapping quality

MLV: murine leukemia virus

MMTV: murine mammary tumor virus

NC: nucleocapsid

NCI: National Cancer Institute

Nf-κB: nuclear factor kappa -light-chain-enhancer of activated B cells

NFAT: nuclear factor of activated T-cells

NIAID: National Institute of Allergy and Infectious Diseases

NIH: National Institutes of Health

NK: natural killer

OCT4: octamer-binding transcription factor 4

ORF: open reading frame

PacBio: Pacific Biosciences

PBMC: peripheral blood mononuclear cell

PBS: phosphate-buffered saline

Pen-Strep: penicillin streptomycin

PERV: porcine endogenous retrovirus

PHA: phytohaemagglutinin

PLZF: promyelocytic leukemia zinc finger protein

PMA: phorbol 12-myristate 13-acetate

PPIA: peptidyl prolyl isomerase A

PR: protease

PRODH: proline dehydrogenase

Psi (ψ): packaging signal

qPCR: quantitative polymerase chain reaction

R: repeat region of LTR

RcRE: Rec response element

RIG-I: retinoic acid inducible gene 1

RLU: relative light unit

RPL: ribosomal protein L

RT: reverse transcription

RT+/-: reverse transcriptase positive/negative

RNASeq: ribonucleic acid sequencing

SA: splice acceptor

SD: splice donor

SETDB1: SET domain, bifurcated 1

SGS: single genome sequencing

SINE: short interspersed nuclear element

SIV: simian immunodeficiency virus

SIVcpz: simian immunodeficiency virus chimpanzee

Solo LTR: solitary long terminal repeat

SP: specificity protein

SRSF1: serine/arginine-rich splicing factor 1

SSBP1: single-stranded DNA-binding protein 1

STLV: simian T-lymphotropic virus

SU: surface subunit of Envelope

SV40: simian virus 40

TBS: tris-buffered saline

TE: transposable element

TM: transmembrane subunit of Envelope

TMC: Tufts Medical Center

TRIM: tripartite motif-containing protein

Tris-HCl: Tris-hydroxychloride

U3: unique region 3' of LTR

U5: unique region 5' of LTR

UCSC: University of California Santa Cruz

UPitt: University of Pittsburgh

XMRV: xenotropic murine leukemia virus-related virus

YY1: yin yang 1

ZNF: zinc finger protein

## Chapter 1: Introduction

Portions of this chapter were previously published in:

Bhardwaj N, Coffin JM. "Endogenous retroviruses and human cancer: is there anything to the rumors?" *Cell Host Microbe*. 2014 Mar; 12;15(3):255-9.

### *1.1 Retroviruses*

*Retroviridae* encompasses a diverse family of enveloped viruses, categorized into seven genera that are related through their unique replication cycle. Initially, retroviruses were termed "oncoviruses" or "RNA tumor viruses" for their well-established ability to cause the cell-free transmission of cancers in chickens and mice [17, 66, 197]. It was decades later that they were hypothesized to replicate their RNA genome through a DNA intermediate [246], a theory that was subsequently confirmed in 1970 [5, 248], which, when discovered, broke one of the central tenants of molecular biology about the flow of genetic information [50]. Specifically, retroviruses have the ability to convert their non-segmented, single-stranded, pseudo-diploid positive sense RNA genome into linear dsDNA through the use of the enzyme reverse transcriptase (RT). Reverse transcription has been recognized to occur in other retroelements and viruses since this seminal finding [20]. After reverse transcription, the retroviral dsDNA genome is then irreversibly integrated into the host genome by the enzyme integrase (IN) as part of its replication cycle [247], a process which in many ways informs the pathology of retroviral infection.

The seven genera of retroviruses are capable of infecting most vertebrate species and are categorized based on the phylogeny of their RT genes [260]. These genera are:

*Alpharetrovirus*, *Betaretrovirus*, *Gammaretrovirus*, *Deltaretrovirus*, *Epsilonretrovirus*, and *Lentivirus*, which are included in the sub-family *Orthoretrovirinae*, and *Spumavirus,* which is a genus in the sub-family *Spumaretrovirinae*. In addition to differences in RT genes, retroviruses vary in genome size, ranging from 7-12kb in length, and can be categorized as either simple or complex retroviruses [39]. Simple retroviruses encode essential elements and genes carried by all members of *Retroviridae*, which notably include *gag*, *pro*, *pol* and *env* flanked by 5' and 3' long terminal repeats (LTRs) in their integrated form, and may include an accessory gene [260]. On the other hand, complex retroviruses, which include members of the *Deltaretrovirus*, *Lentivirus* and *Spumavirus* groups, encode multiple accessory genes in addition to all essential genes, which are necessary for regulation of gene expression and important for control of the host response in many cases [260].

If we consider the simple retrovirus as the standard, there are two major transcripts produced during infection: (1) a genomic RNA which contains all the aforementioned essential genes; and (2) a subgenomic transcript containing only the *env* open reading frame (ORF) (Figure 1-1 A) [192]. In addition, note that complex retroviruses and even some simple retroviruses carry multiple splice donor and acceptor sites and therefore produce multiply spliced subgenomic transcripts encoding accessory genes, however these are variable. Commonly, these transcripts are produced by the host RNA polymerase II through its recognition of cis-acting elements on the integrated retroviral 5' LTR, but transcription may be enhanced by the presence of trans-activating accessory proteins encoded by the retrovirus [192]. Two copies of the genomic transcript are preferentially packaged into nascent virions through the presence of a packaging signal

(ψ), which is a secondary structure formed by the presence of specific sequence, usually in the 5' leader of *gag*, on the viral genomic RNA [237]. The ψ signal enhances viral genome packaging over cellular genes 20-200-fold [237].

The genomic RNA carries the essential genes *gag-pro-pol-env*, which can be packaged into virions or translated into Gag, Gag-Pro and Gag-Pro-Pol polyprotein [237]. Based on genomic structure (Figure 1-1 A), *gag*, *pro* and *pol* ORFs may minutely overlap or be in-frame, and can be translated off the same genomic transcript [237]. However, since the *gag, pro* and *pol* genes are not all in-frame at the same time, Gag, Gag-Pro and Gag-Pro-Pol polyprotein production is dependent upon translational read-through of the termination codon after *gag* into the *pro* or *pol* ORFs and/or -1 frame shifting events that allow for ribosomal recognition of the downstream ORFs [237]. Due to the complexities of translation of the downstream ORFs, Gag protein is commonly seen at levels 10-20x higher than Gag-Pro and Gag-Pro-Pol [237]. The *env* ORF is not translated off the genomic transcript, but rather from a dedicated subgenomic transcript (Figure 1-1 A) [237].

The *gag* gene encodes the Gag polyprotein that contains the structural genes necessary for retroviral replication, which includes matrix (MA), capsid (CA) and nucleocapsid (NC), as well as minor phosphoprotein and/or peptide products (Figure 1-1) [237]. Gag polyprotein is responsible for driving retroviral particle assembly, as virions have been produced with only Gag expression in the cell [237]. MA is important for membrane binding of Gag and Env glycoprotein incorporation into virions; in addition, MA forms a lattice under the membrane of the mature viral particle [237, 244]. CA is needed for particle assembly and makes up the capsid core that surrounds the viral

genome and enzymes in a mature particle [79]. NC binds retroviral genomic RNA primarily through use of basic residues [38], and facilitates the multimerization of Gag polyprotein at the cell surface [171].

Protease (PR), encoded by *pro*, is one of the essential enzymes of the retrovirus [260]. PR is necessary for cleaving the Gag, Gag-Pro and/or Gag-Pro-Pol polyprotein in an immature virion into the mature forms of Gag (MA, CA and NC) and Pol (RT and IN), which leads to dramatic morphological changes and maturation of the viral particle [260]. PR is active only after dimerization, which appears to occur late in assembly or after budding of the viral particle [260].

The *pol* gene encodes genome replication-associated enzymes, which are the aforementioned RT and IN proteins that drive the unique retroviral lifecycle. RT, which is the most conserved gene between genera, is the driver of reverse transcription, where it utilizes its three functions as an RNA-dependent DNA polymerase, DNA-dependent DNA polymerase and RNase H. In reverse transcription, a dsDNA copy of the virus is created from the viral RNA genome using a tRNA molecule as a primer for first strand synthesis; during this process, the viral RNA genome is degraded by the inherent RNase activity of RT when in duplex with nascent viral DNA [260]. Reverse transcriptase has a high error rate (~$10^{-4}$ -$10^{-5}$ errors/base) and is able to template switch between the two packaged viral RNA genomes at sites of homology to make a recombinant dsDNA [245], both features that are important to the observed diversity of retroviruses. The produced viral dsDNA is part of a pre-integration complex (PIC) that is targeted to the cell nucleus, mediated by the presence of retroviral CA in the PIC [157], and the retroviral dsDNA is integrated into the host genome through the function of IN. At this point, the integrated

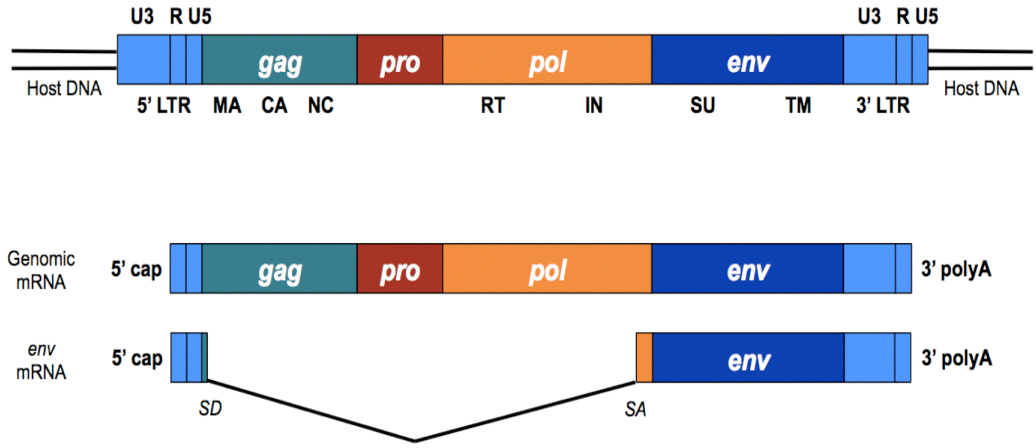
copy of the retrovirus is referred to as a provirus and remains as an irreversible insertion. Dependent upon the retrovirus, IN may be targeted to certain regions of the genome, for example, near transcriptional start sites [273] or areas of active transcription [218], based on interaction with host factors, however, in some cases, integration is completely random [269]. Regardless of the mechanism, the site of proviral integration has important implications for retroviral pathogenesis.

The *env* ORF encodes the glycoprotein Env, which studs the membrane of the retroviral particle. Env is present as a trimer and is made up of two main regions, the surface (SU) region, which mediates binding to the cognate receptor, and the transmembrane (TM) region, which contains the fusogenic machinery that mediates virus entry [104]. Env is initially a single precursor polypeptide that is trafficked from the ER to the Golgi, where it is glycosylated and cleaved by a host protease (by furin or similar enzyme) into its two domains SU and TM, which remain associated by non-covalent interactions or disulfide bonds [260]. Env further traffics to the cell surface where it co-localizes with the nascent virion.

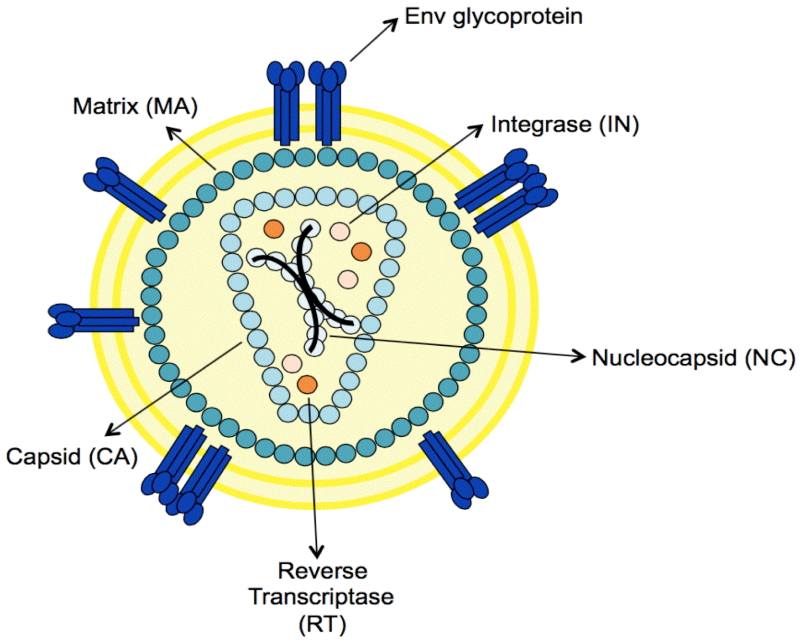
Surrounding the *gag*, *pro*, *pol* and *env* genes on the provirus are two LTRs. Due to the requirements of reverse transcription, at the time of retroviral integration the 5' and 3' LTRs are identical and comprise three regions: unique region 3' (U3), repeat (R) and unique region 5' (U5). The U3 region contains enhancer sites for transcription factor or hormone binding, which vary dependent on the retrovirus, and also a promoter site at the U3-R boundary, which has a TATA box and other elements [192]. The R region defines the transcriptional start for the retrovirus and also serves as the site of 3' end processing [192]. As there are both 5' and 3' LTRs, a provirus can drive its own transcription as well

**Figure 1-1. A retrovirus genome (A), virion (B) and replication cycle (C).**

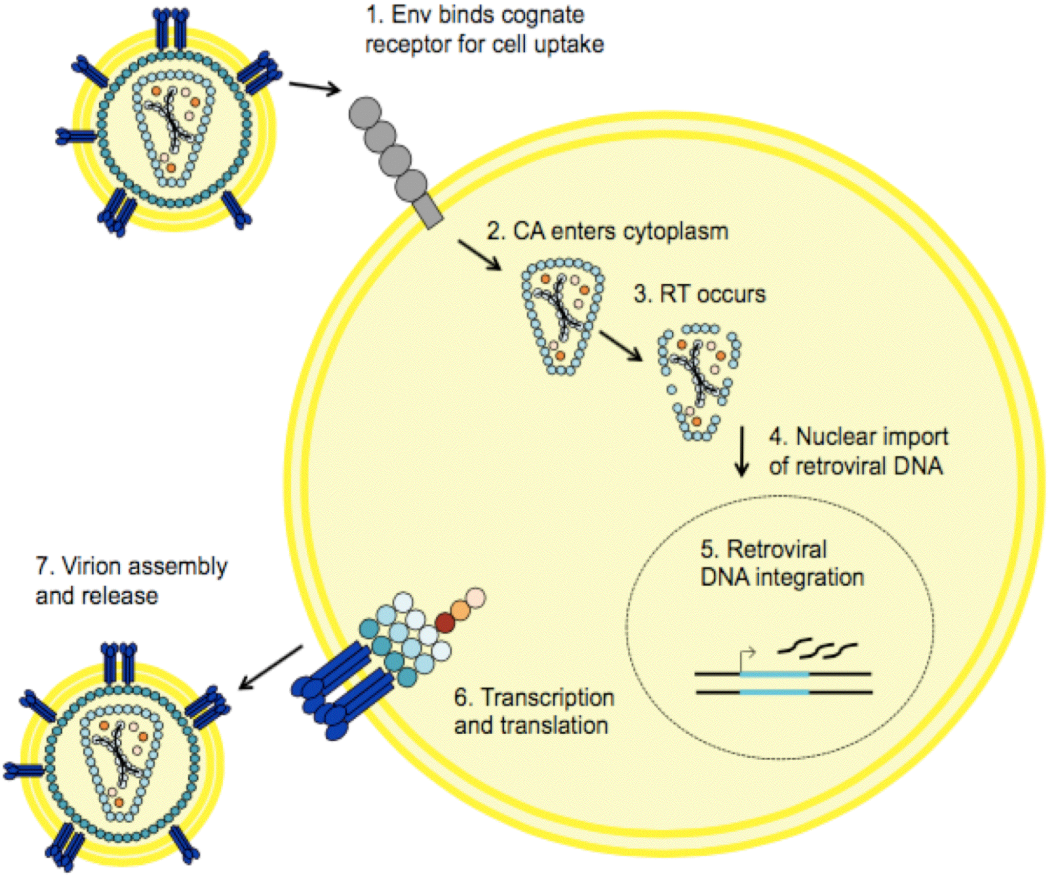
A




B



C


1. Env binds cognate receptor for cell uptake
2. CA enters cytoplasm
3. RT occurs
4. Nuclear import of retroviral DNA
5. Retroviral DNA integration
6. Transcription and translation
7. Virion assembly and release

as the transcription of downstream host genes, a property that relates directly to its potential pathogenicity in a host.

*1.2 Retroviral oncogenesis*

The defining moments of the retrovirus lifecycle, reverse transcription and integration, inform their pathogenesis in animals. Many retroviruses are capable of causing cancers, and in fact, this is how they were initially discovered. Starting with the process of retroviral integration, the presence of the provirus and its promoter elements can disrupt cellular regulation of oncogenes in a process called proviral insertional mutagenesis [196]. If a provirus is randomly inserted near a cellular oncogene, its regulatory elements can lead to increased recruitment of transcriptional machinery to the gene or even lead to 5' or 3' LTR-driven transcription of the oncogene from an upstream inserted provirus. Dysregulation of oncogenes can lead to oncogenesis; for example, in chickens, this exact mechanism is at play when the alpharetrovirus avian leukosis virus (ALV) was found to integrate near *c-myc* in bursal lymphomas and drive transcription from its 3' LTR [96]. If instead expression is driven from the retroviral 5' LTR, there is the potential that a genomic transcript will contain read-through sequence of a cellular oncogene and be packaged into budding virions. Carriage of an oncogene has been shown to occur with many retroviruses, including avian, feline and mouse retroviruses [196], and leads to quickly transforming retroviral infection [236], though usually in the presence of a helper virus.

Reverse transcription can also affect the ability of a retrovirus to cause cancer in a host. As previously mentioned, a viral particle packages two copies of its viral genome,

which is produced from proviral transcription of genomic RNA. However, the integrated copies of a retrovirus may not be infectious or oncogenic due to mutations in the viral genome. If two defective yet complementing copies of a viral genome are expressed within the same cell and packaged in the same virion, after infection of a new cell, template switching during reverse transcription can lead to the production of a recombinant replication-competent virus.

Recombination leading to oncogenesis has been observed in animal models of infection. In the AKR mouse model of oncogenesis, a leukemogenic retrovirus is generated from a coordinated series of recombinations between three germ-line encoded murine leukemia virus (MLV) elements, which donate different parts of their genomic sequence (e.g. LTR or polytropic *env* sequence capable of infecting murine and non-murine cells), to create an infectious oncogenic virus, but do not cause cancer individually [230]. In another example, in the setting of immunodeficiency, replication of germ-line encoded MLV was shown to increase in $Rag1^{-/-}$ mice, likely due to microbial stimulation in the gut; here, increased expression led to recombination between endogenous MLVs to create an infectious virus capable of causing lymphomas [278]. Thus, the position of proviral integration and likelihood for recombination are crucial in assessing the oncogenic potential of a retrovirus.

*1.3 Exogenous human retroviruses*

The first retrovirus discovered to cause human disease was human T-lymphotrophic virus (HTLV-1) in the late 1970s [190, 254]. HTLV-1 is a deltaretrovirus and is estimated to infect 5-20 million people worldwide [55], though its geographic distribution

is very irregular, with HTLV-1 endemic populations clustered in southwestern Japan, the Caribbean, Papua New Guinea and parts of Africa, the Middle East and South America [84]. It is thought to have evolved from simian T-lymphotrophic virus (STLV-1) simian-to-human transmissions between 3,000-21,100 years ago depending upon the subtype of HTLV-1, with separate STLV-1 transmissions causing multiple HTLV-1 subtypes, though interestingly the prevalent HTLV-1 subtype A did not appear to cluster with any STLV-1 isolates in phylogenetic analyses [123, 257]. HTLV-1 preferentially infects CD4+ T cells and disease is mostly asymptomatic, though a portion of infected individuals progress to neurological or malignant disease, the pathogenic outcome that prompted HTLV-1 discovery [83, 258].

Shortly after the discovery of HTLV-1, human immunodeficiency virus type 1 (HIV-1), the causative agent of acquired immunodeficiency syndrome (AIDS), was recognized [69]. AIDS was identified in 1981 [33, 34], and was found to be caused by the lentivirus HIV-1 in 1983 [8, 78]. The resultant HIV pandemic has infected over 60 million people worldwide, with a death toll of over 25 million [61], and is primarily caused by infection with HIV-1 group M [223]. HIV-1 group M originated from simian immunodefiency virus chimpanzee (SIVcpz) that jumped into humans and was back calculated to have existed in humans since the 1920s [80, 124]. Similar to HTLV-1, HIV-1 infects CD4+ T lymphocytes, though infection with HIV-1 is characterized by systemic immunodeficiency due to their depletion [40, 53, 69]. The dynamic cycle of infection and clearance of infected cells culminates in the destruction of this immune cell population, either through the direct effects of infection, immune surveillance or apoptosis/pyroptosis induced by increased cellular activation [40, 62, 69, 71]. HIV-1 immunosuppression

leads to increased susceptibility to opportunistic infections, like *Pneumocystis jirovecii*, *Candida albicans* or *Mycobacterium tuberculosis*, and fatal neoplastic diseases as the critical CD4+ T cell subsets are exhausted [69, 217]. While there are antiretroviral medications available that can effectively suppress viral replication with proper adherence [4], these interventions are not equally available to all infected individuals. Furthermore, there is no known cure for HIV-1 infection [19] and there are associated cardiovascular and age-related comorbidities [31, 252], likely related to elevated levels of immune activation and inflammation in infected individuals.

These two circulating human retroviruses are known to lead to cancer, though by differing mechanisms. HTLV-1 causes adult T-cell leukemia/lymphoma (ATL) in a minority of infected patients (3-5%) [106]. ATL is thought to be caused by the activities of HTLV-1 accessory proteins Tax, which activates Nf-κB and Akt signaling and inhibits p53 function, and HBZ, which supports the proliferation of ATL cells [106, 281]. Currently, there is no clinical test available to predict which patients will progress to ATL, though they are often adults infected for 20-30 years, and the clinical course of ATL is heterogenous [106]. HIV progresses to malignancy in patients due to immune deficiency and opportunistic infections. Many HIV-associated cancers are known to arise by co-infection with another virus, for example human herpesvirus-8 (HHV-8) in Kaposi's sarcoma and primary effusion lymphoma, and Epstein-Barr virus (EBV) in primary central nervous system lymphoma [24, 32, 162, 212, 276]. Interestingly, HIV has not been shown to transform infected cells through delivery of a transforming protein or via insertional mutagenesis. Direct oncogenic effects of HIV-1 infection due to insertional effects have been proposed [225] and in fact recent results show that HIV-1

provirus integration may drive clonal expansion of infected cells *in vivo* [151], suggesting that this possibility should be further investigated. Therefore, oncogenesis caused directly by the retrovirus or due to its disease manifestations are relevant to human retroviral infections.

### *1.4 ERVs: The retroviruses in our genome*

The genomes of all mammals, and indeed of most or all vertebrates and many invertebrates, contain elements known as endogenous retroviruses (ERVs), which are recognized to be more than just "junk DNA." Though infection generally occurs through horizontal transfer, where a retrovirus infects a somatic cell, replicates and is passed from cell to cell and from one individual to another (e.g. HIV and CD4+ T cells), retroviruses can also be inherited and transferred genetically, from parent to child, due to their irreversible integration into host cell DNA. These elements that are passed vertically in humans, the human endogenous retroviruses (HERVs), are the descendants of infection of the germline cells of our ancestors, which is known to have been occurring over 100 million years [109, 133]. Almost half of the human genome is composed of parasitic elements, which includes DNA transposons (2.8%), non-LTR retroelements (33.9%) and LTR retroelements that include HERV sequences (8.3%) [126]. There are over 30 families of HERVs known today that have populated the human genome either through serial infection of germ line cells as previously mentioned and, for a few families, through retrotransposition, where retroviral DNA is transcribed into RNA, reverse transcribed into DNA and integrated into the genome of the same cell, lacking an extracellular infection phase [11]. Once integrated into the genome, HERVs are inherited

in a Mendelian fashion, akin to genes, and subject to similar selection pressures, as insertions can have beneficial, detrimental or neutral effects on the host [109].

*1.5 Function of ERVs in the genome*

As previously mentioned, at the time of integration, the inserted retroviral genome, called a provirus, contains all essential elements needed for its replication. The provirus contains the 5' and 3' LTRs, which contain promoter, enhancer and polyadenylation sites, which surround the canonical retroviral genes *gag*, *pro, pol* and *env*. The provirus can be deleterious to the host for multiple reasons. Most evidently, a provirus can encode a replicating virus. Even though its expression may be strongly suppressed, it can lead to infection of the whole animal, with pathogenic consequences such as cancer due to insertional activation of gene expression, immunodeficiency, and others. Expression of individual retroviral genes can also have ill effects, such as cell-cell fusion mediated by expression of a fusogenic Env glycoprotein, which has been implicated in ovarian cancer [264]. The provirus, depending upon where it inserts, can also influence the expression of neighboring genes through the use of regulatory sequences present on the LTRs or splice donor and acceptor sites present in the genomic sequence [70]. In addition, as retroviruses have been infecting the germline for over 100 million years, a provirus can be used as a site of homologous recombination with established ERVs and cause genomic rearrangements [103]. In short, although HERVs can be benign, or even beneficial, retrovirus infection presents a special threat to the host, not only for the infected individual, but also for descendants who may inherit the HERVs resulting from that infection.

There is substantial pressure to minimize the effects of retrovirus infection on the host. The relative contributions of somatic infection and HERV formation to this pressure are unclear. HERV insertions that are highly detrimental to the host are unlikely to be fixed in a population under conditions of normal gene flow (i.e. no bottleneck events). It seems plausible that if initially detrimental insertions are fixed in a population, they would have been inactivated prior to fixation [6]. For those insertions that reach fixation, including HERVs that were defective upon integration or integrated into non-critical regions of the genome, most have accumulated mutations that have rendered them non-functional or non-infectious to the host species. In the human genome, HERVs are underrepresented in gene-rich areas and are more likely to be found in intergenic regions [159], an effect that increases with age of the provirus and attributable to counterselection of those integrants most likely to affect gene expression [109]. Furthermore, even though HERVs comprise ~8% of the human genome, none is known to be infectious [109]. The survival advantage for non-infectious ERVs in genomes was recently recognized. Groups of "*env*-less" ERVs expanded up to 30 times more than ERVs maintaining *env* in mammalian genomes, where expansion was achieved through retrotransposition [149]. Though these elements retain their capacity for insertional mutagenesis, this observation implies an expansion advantage for ERVs that sacrifice their extracellular lifecycle, potentially due to the less harmful effects of retrotransposition on host fitness as compared to reinfection cycles.

Beyond mutational changes, inactivation of HERVs post-integration also results from formation of solo LTRs, which are formed when the proviral 5' and 3' LTRs, which are identical at integration, recombine and excise the proviral genome, leaving only one LTR

sequence in place of the full provirus [231]. While this event eliminates the retroviral genome, the LTR sequence can still provide a site for recombination or affect the expression of nearby genes [109]. For example, a hypomethylated LTR derived from a retrotransposon was found to direct transcription of the proto-oncogene colony-stimulating factor 1 (CSF1R) in Hodgkin's lymphoma, necessary for survival of the malignant cell [128]. Though the exact ratios differ by group of HERVs, solo LTRs can outnumber proviruses 10:1, another way in which the remains of replication-defective proviruses are maintained in the genome [231].

The series of retroviral insults on the mammalian ancestral lineage may have contributed to the development of innate immune defenses to prevent further infection. These defenses include the APOBEC family of proteins [95, 209], which are cytidine deaminases that can mutate the nascent retroviral DNA during RT, and are present only in one copy in rodents but expanded in primates [108]. Another factor, Trim5α, restricts retroviral replication by binding CA, which is released into the cytoplasm after fusion with host cellular membrane [210, 232]. Interestingly, some retroviral restriction factors emerged from ERVs themselves. Two ERVs present in mice have evolved to become restriction factors, known as Fv1 and Fv4. Fv1, derived from the *gag* region of a spumavirus-like endogenous element in the mouse genome, restricts an unrelated MLV by binding the CA lattice and halting nuclear import of the retroviral complex, reminiscent of Trim5α restriction [12, 207]. Fv4, derived from a murine ERV *env* gene, blocks receptors for exogenous MLV and prevents new infection in a manner akin to super-infection resistance [175]. The phenomenon of receptor interference was first observed with ALV susceptibility, where genetically dominant traits, discovered to be

Env glycoproteins from proviruses, provided resistance to infection [187, 195]. In fact, other mammals currently being infected by exogenous retroviruses use endogenized elements to similar effect. In sheep, expression of Env derived from endogenous Jaagsiekte sheep retrovirus (JSRV) binds receptors needed by exogenous JSRV to enter cells, thus providing protection from lung cancer caused by JSRV infection [183].

Co-opting ERVs can provide an avenue for novel retroviral restriction mechanisms, as explained above, but also for new physiological functions relevant to host evolution. The most impressive examples of preserved HERV functionality are the use of *syncytin-1* (HERV-W; on chromosomal band 7q21.2) and *syncytin-2* (HERV-FRD; on chromosomal band 6p24.1) in placentation, both of which are ancient human ERV *env* genes utilized for the fusion of cytotrophoblasts to syncytiotrophoblasts due to their fusogenic (*syncytin-1, syncytin-2*) and immunosuppressive properties (*syncytin-2*) [18, 163]. In a noteworthy example of convergent evolution, the process of co-opting endogenous retroviral Env for use in placentation has occurred independently in multiple lineages of eutherian mammals [49, 65, 182].

In addition to ERV genes, ERV LTRs can be co-opted for uses beneficial to the host. For example, a HERV element integrated upstream of the pancreatic amylase gene allows for its expression in the salivary glands, a function that permits humans to digest starches in the mouth [205]. In another example of LTR function, one-third of the sites bound by the tumor-suppressor p53 in humans are ERV LTRs, the result of integrations from >40 mya [262]. Potentially, ERV LTRs with functional motifs can be used to expand gene networks responsive to cell physiology, where the LTR motifs could allow for cell-specific gene expression or activation in response to cell stress. LTRs can also drive the

expression of regulatory sequences like long intergenic non-coding RNAs (lincRNA) [117]. In fact, 10% of human lincRNAs were found to be driven by an ERV LTR, and 75% of all lincRNAs contain sequence derived from a transposable element (TE), implicating TEs (and specifically ERVs) in the evolution and diversification of lincRNAs [117]. Impressively, in humans, hypomethylated HERV-H LTRs drive the expression of important lincRNAs (like *linc Regulator of Reprogramming (RoR)* [140]) and recruit pluripotency factors and transcriptional activators like OCT4 and p300 to the genome, both crucial for maintenance or reprogramming of a pluripotent state [145, 176].

*1.6 Epigenetic control of ERV expression*

Although expression of some ERVs can be of benefit, uncontrolled expression of retroviral gene products would not be expected to have beneficial effects for the host, especially for newer, active integrations. Much knowledge on ERV control has been derived from detailed studies in mice. The study of mouse embryonic stem cells led to the discovery of a complex genomic surveillance system that causes transcriptional repression of murine ERV elements through epigenetic regulation. In embryonic cells, specific KRAB zinc finger (ZNF) proteins, capable of recognizing varied DNA motifs, appear to be able to target the primer-binding sites of ERVs that utilize a proline tRNA to prime RT, a group that includes coding competent MLVs as well as VL30 elements, which do not contain intact ORFs [272]. KRAB-ZNF binding then recruits the co-repressor scaffold protein KAP1 (also known as TRIM28) [198, 270], which is able to recruit the histone methyltransferases SETDB1 [219] and ESET [158] that deposit repressive histone 3 lysine 9 (H3K9) methlyation marks, and the heterochromatin protein

HP1 [271], among others, to create a microenvironment of repression. In addition, yin yang 1 (YY1) transcription factor binding sites present on ERV LTRs were shown to serve as sites for KAP1/TRIM28 recruitment [214]. Deletion of the three major DNA methyltransferases did not significantly affect embryonic ERV repression in mice, as measured by reactivation of MLV and the *env*-less IAP and MusD elements, implying a larger role of histone modification for ERV silencing during embryogenesis over DNA modifications [158]. Interestingly, the number and age of the KRAB-ZNF/KAP1 surveillance system genes in mammals is associated with the rise in number of ERVs in their genomes [67]. Outside of the KRAB-ZNF system, gene therapy experiments led to the discovery that hypoacetylated histone 3 (H3) and bound histone 1 (H1) are associated with silenced retroviral and lentiviral insertions in mouse embryonic stem cells; in these experiments, DNA methylation had a role in transcriptional repression but was not required [275].

CpG methylation of DNA can control ERV expression [277], as seen with increased expression in differentiated tissues after treatment with the demethylating agent 5-azacytosine [91, 99]. Methylation at cytosine residues occurs *de novo* with the enzymes DNMT3a and DNMT3b, generally during embryogenesis, and the methylation marks they deposit are maintained through the activity of DNMT1 in differentiated tissues [199]. In contrast to embryonic cells, the deletion of KAP1/TRIM28 in a fibroblast cell line did not increase ERV expression in mice, illustrating the importance of DNA modification to control of ERV expression in differentiated tissues [198].

The relevance of these mechanisms to the control of human integrations is not well documented and requires substantially more research. Contrary to the control of

MLVs and other murine ERV elements, the activity of HERV LTRs in human placenta appeared to be related to their DNA hypomethylation state and genomic context [193]. Recently, a complex of three proteins called the HUSH complex was shown to interact with SETDB1 to deposit repressive H3K9 trimethylation marks to heterochromatic regions of the human genome to maintain transcriptional repression in differentiated cells, independent of KAP1/TRIM28 [243]. *In vitro*, the HUSH complex was reported to repress HIV and MLV proviruses located in heterochromatin, suggesting that this complex has potential to act on ERV integrations as well. In addition, in differentiated human tissues, there appear to be large gene-poor regions (up to 5Mb) of heterochromatin bounded by H3K9 dimethylation, deposited by the histone methyltransferase G9a, that encompass ~30% of the genome, dependent on the tissue type [268]. These regions, called large organized chromatin K9-modifications (LOCKs), could play a role in modulating ERV expression. In general, however, there is a poor understanding of HERV expression patterns in human tissues and how they are regulated. Due to the diversity of HERVs, it is likely that the regulation of different HERV groups may differ based on their potential for expression and positions in the genome. The relative contributions of higher-level epigenetic control and DNA methylation in embryonic and differentiated human tissues remain to be elucidated.

### *1.7 The HML-2 group of HERVs*

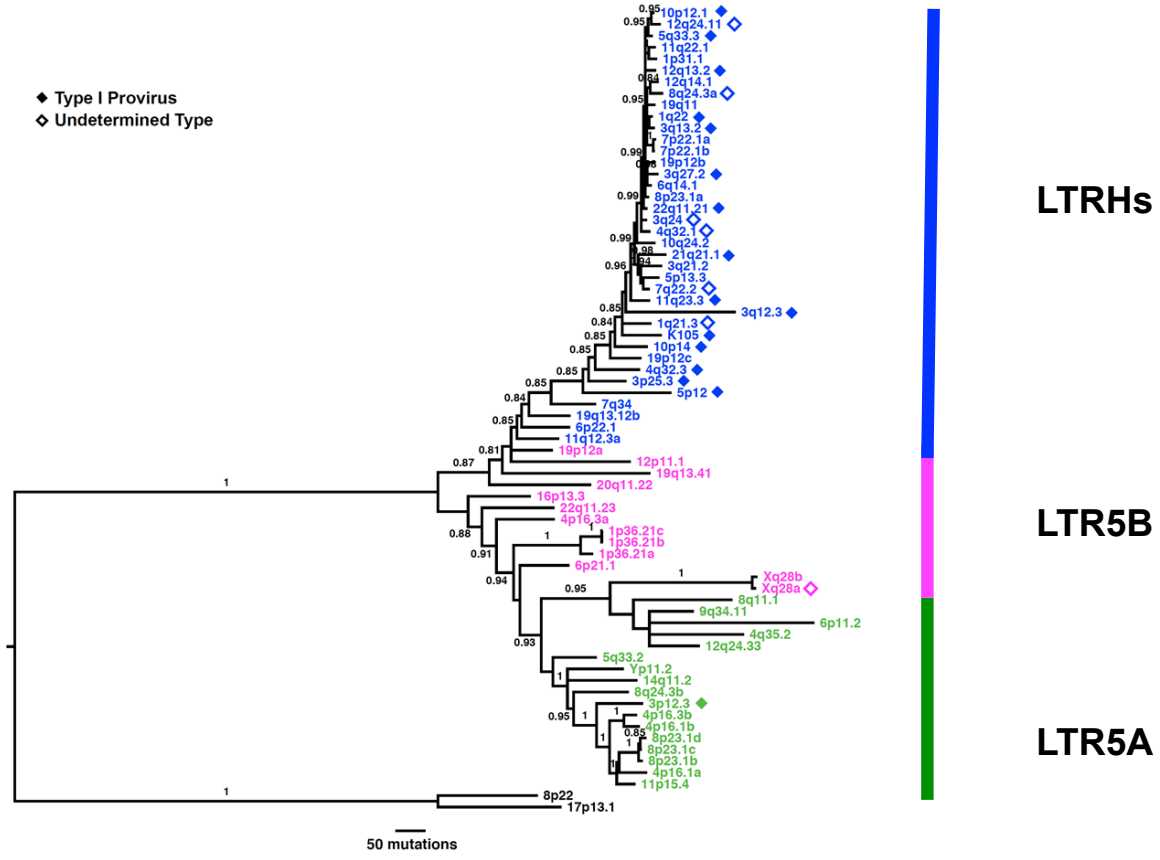
Many HERV sequences have undergone extensive mutation after integration into the genome; however, a few have maintained ORFs for viral proteins and have the ability to form intact but noninfectious, viral particles [6, 23, 59, 144, 153]. One group of HERVs,

named HERV-K (to reflect the use of lysine tRNA to prime reverse transcription), includes many proviruses that have retained ORFs [144, 233]. This group is made up of 11 subgroups named to reflect their similarity to mouse mammary tumor virus (MMTV), a betaretrovirus that exists in exogenous and endogenous forms and causes mammary carcinoma in mice [6, 17, 107, 233]. The human HERV-K MMTV-like (HML)-2 group began infecting the ancestral human germline ~30-35 million years ago, after the New World and Old World monkey divergence [228]. Interestingly, the HML-2 group includes proviruses that have most recently integrated into the genome and are evolutionarily young compared to all other HERV sequences, with some having entered our genome soon after the Human-Chimpanzee divergence and up to within the last few hundred thousand years [10, 111].

Due to their recent infection of the germ line, HML-2 proviruses have unique qualities when compared to other HERVs present in the genome. This is the only subgroup to include human specific integration sites, where at least 11 are polymorphic within the human population [7, 10, 102, 233, 253]. Furthermore, this subgroup contributes >90 full- or near full-length proviral sequences to the human genome, many of which contain an ORF for *gag*, *pro*, *pol* and/or *env,* genes essential for infectivity [233], as well as over 950 solo LTR elements. Three types of HML-2 proviruses have been categorized based on their LTR phylogenies and are referred to as LTR 5B, LTR 5A and LTR Hs HML-2 proviruses [233]. LTR 5B proviruses are basal to both LTR 5A and LTR Hs proviruses. Of interest, the LTR Hs proviruses include the recent, human-specific HML-2 integrations [233]. The HML-2 LTR contains binding sites for transcriptional factors that have been experimentally shown to stimulate its activity,

**Figure 1-2. Gag phylogeny of HML-2 proviruses displaying LTR subtypes.**

(**A**dapted from [233])




which includes the ubiquitous cellular factors Sp1, Sp3 and YY1, the immune associated factors Nf-κB and NFAT, and hormone binding sequences for estrogen, progesterone and androgen [154].

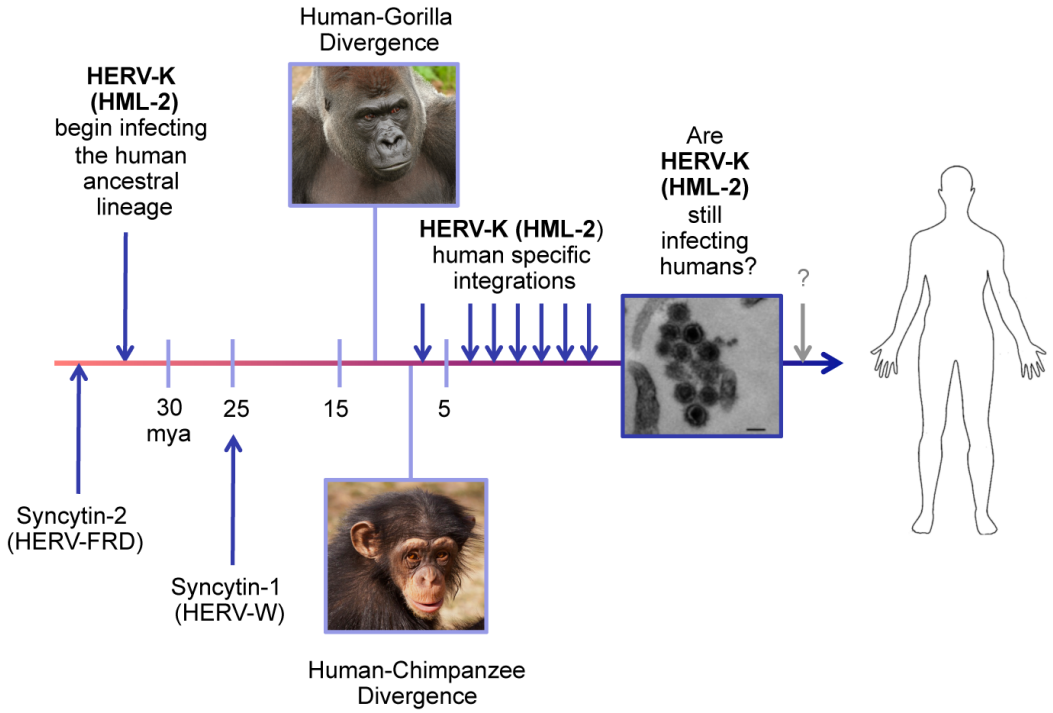
The relatively recent entry of HML-2 proviruses into the human genome and their sequence preservation has led to speculation about whether any of these proviruses can encode an infectious retrovirus. The provirus K113 (on chromosome 19p12) retains full coding capacity for all retroviral genes; however, it was not found to be replication competent in vitro [9]. Two groups have engineered recombinant HML-2 viruses based

on the most recent common ancestor of the human-specific proviruses, as inferred by phylogenetic analysis, that have been shown to be weakly infectious [60, 134]. Based on these findings, recombination between as few as three HML-2 proviruses could lead to the production of an infectious retrovirus [60]. Recombination is a frequent event in reverse transcription and has been shown to occur in multiple animal models; recently, the phenomenon was shown to have produced xenotropic MLV-related virus (XMRV) [186], a mouse-derived gammaretrovirus alleged to be the causative agent of chronic fatigue syndrome in humans, though these claims were unsubstantiated [141, 226]. The recombinant HML-2 virus was restricted by APOBEC3F but not APOBEC3G or Trim5α [134]; however, ancestral HML-2 integrations show signs of APOBEC3G mutation [135].

Though unproven, it remains possible that an exogenous form of HML-2 is still infecting the human germline, occurring at a low frequency within the population. In fact, due to the low frequency of polymorphic integrations, some HML-2 insertions are not annotated in the human genome builds [146, 233]. Multiple groups have discovered HML-2 integrations that are unique to humans [156] and a few which may even be unique to the Neandertal and Denisovan lineages [1, 155]. In support of their potential pathogenicity to humans, one of the reconstituted HML-2 viruses was shown to have a preference for integration into active transcriptional units [25]. In our genome, ERV integrations are mostly present outside of active genes [159], a sign of the intense negative selection against retaining integrations with pathogenic capacity. By this logic, recent integrations have the greatest potential for disease.

**Figure 1-3. Integration of HML-2 viruses over evolutionary history.**

(Adapted from [13])




Of interest, HML-2 proviruses can be categorized as type 1 or type 2 proviruses [6]. This distinction refers to a characteristic 292-bp deletion in the SU region of the *env* gene of type 1 proviruses, which is associated with the production of a presumed alternative accessory protein named Np9 [2], whereas type 2 proviruses that retain full *env* coding sequence produce the accessory protein Rec [143]. Rec is functionally analogous to an accessory protein encoded by HIV-1 called Rev [143, 152, 274]. Rec shuttles unspliced or partially spliced HML-2 mRNA out of the nucleus into the cytoplasm by binding an RNA secondary structure on the U3 region of the LTR, called the Rec response element

(RcRE) [147, 148]. Conversely, Np9 has no known function in the HML-2 replication cycle.

The two accessory proteins Rec and Np9 are unique to the HML-2 group of HERVs and are not associated with a known step in human development or biology; moreover, they are considered putative oncogenes [126, 203]. Rec, when expressed in a transgenic mouse model, led to the disruption of germ cell development and caused changes similar to carcinoma in situ, which are the precursor lesions of germ cell tumors [77]. Both Rec and Np9 are reported to bind and disrupt function of the tumor suppressor promyelocytic zinc finger protein (PLZF), though they target different regions, which normally acts as a transcriptional repressor of the proto-oncogene *c-myc* [57, 119]. Additionally, Np9 was reported to bind the E3-ubiquitin ligase called ligand of Numb protein X (LNX), which, as its name implies, binds Numb, the negative regulator of the proliferative Notch pathway [3]. This interaction is hypothesized to lead to undue proliferative signals, though conclusive evidence of this was not observed. Though both *rec* and *np9* expression are associated with proliferation and oncogenesis, transcripts from both of these accessory genes have been reported in healthy tissues to no apparent effect [216].

### *1.8 HML-2 association with human disease*

The similarity of HML-2 to MMTV and its capacity for mobilization has prompted steady research into its pathogenic potential, especially in terms of carcinogenesis. In placental, embryonic and malignant tissues, HML-2 expression as RNA, protein and even intact but non-infectious particles has been reported [75, 92, 116]. Increased expression could be due to decreased methylation of DNA, altered histone regulation or specific

changes in signaling and transcription factor expression that take place (e.g. steroid hormone expression in breast cancer) [98]. For example, hypomethylation and binding of the transcription factor OCT4 appeared responsible for induction of HML-2 expression in human pluripotent stem cells [92]. Pathogenic effects of HML-2 transcription and protein expression in humans are unknown, though HML-2 expression appeared to protect pluripotent cells from exogenous viral infection [92].

HML-2 was first linked to human disease with the discovery that an HML-2 provirus encodes the human teratocarcinoma derived virus (HTDV) particles emanating from germ cell tumors (GCT) and cell lines [27, 142]. These viral particles are not infectious and indeed, their functional significance is unclear. In studies of GCT patients, reports of HML-2 transcript, Env and Gag protein expression are abundant and antibody responses against these antigens were observed to occur in 50-80% of patients [22, 208]. Remarkably, new types of spliced transcripts encoded by HML-2 were discovered in these cells, the aforementioned accessory genes *rec* and *np9* and a transcript of unknown function named *hel* [143].

Similar to GCT, melanomas and breast cancers are also highly associated with HML-2 expression. In melanoma, the retroviral proteins Gag, Env, RT, Rec and Np9 are produced and approximately 16-22% of patients with melanoma develop an antibody response against HML-2 Env and Gag [28, 94]; interestingly, the antibody response had a negative correlation with patient survival [94]. Retroviral particles have also been observed in melanoma cell lines, where particles from the 518A2 cell line were reported to infect bovine MDBK cells [172]. However, an infectious virus has not been isolated from these cells, and infection was not seen using particles produced from the SKMel28

cell line [28]. An interesting quality of melanomas is that 85% of tumors express a HERV-K protein, a short *env* ORF from the HML-6 provirus named HERV-K-MEL, which can promote specific cytotoxic CD8+ T-cell immune responses [213]. HERV-K-MEL was determined to be a marker of melanoma risk and expression is seen in a few other tumor types [213]. Thus, expression of HERV proteins could potentially be targeted by the immune system as tumor specific antigens, acting as biomarkers instead of (or in addition to) playing a pathogenic role. Likewise, the idea of targeting HERV antigens in tumors is being explored with the HERV-E derived HLA-A11 restricted CT-RCC-1 peptide for treatment of renal cell carcinoma [240].

Breast cancer cells from patients and cell lines similarly show high levels of HML-2 expression in the form of mRNA and protein. Retroviral particles bud from a breast cancer cell line after treatment with estrogen and progesterone [68]. Protein expression has been reported to stimulate both humoral and cytotoxic immune responses in patients [265, 267]. Intriguingly, metastasis to lymph nodes was more likely to occur in tumors expressing HML-2 Env in cancer cell implantation studies in mice [266]. Treatment of breast cancer cell lines with HML-2 Env specific antibodies impede their proliferation and induce apoptosis *in vitro*, as antibody treatment corresponded with overexpression of the p53 pathway, associated with apoptosis and cell senescence [266]. Antibody treatment also reduced the growth of xenografted tumors in mice *in vivo* [266]. The applicability of this finding to human breast cancer and metastasis is still being investigated. In theory, HML-2 Env has preserved both its fusogenicity needed for entry into target cells as well as immunosuppressive activity though exposure of an immunosuppressive domain (ISD) in the TM region [169, 203, 279], though the presence

of an ISD on HML-2 Env has been contested [54]. Therefore, expression could contribute to oncogenesis by fusing neighboring cells and causing genomic instability, or by allowing for tumor immune evasion by expression of this surface glycoprotein. While these properties are readily recognized in other retroviral Env, like syncytin-2, they are not well established for all HML-2 proviruses with *env* ORF.

*1.9 Association of HML-2 expression with HIV-1 infection*

Recent reports indicate a relationship between HML-2 expression and HIV-1 infection, especially intriguing due to the possible dynamics of an exogenous retrovirus influencing the expression of an endogenous retrovirus. Several studies have reported high levels of HML-2 RNA ($10^3$ to $10^{10}$ copies/mL) in the blood of >95% of HIV-1 patients as compared to Hepatitis C infected and/or seronegative controls, which only showed particle expression in 5-8% of subjects [42, 43, 45, 46]. HML-2 particles were reported to co-fractionate with HIV-1 particles in density gradients, though electron microscopy was not able to discern the detailed morphology of the particles [43, 46]. Interestingly, HIV-1 infected patients on suppressive highly active antiretroviral therapy (HAART) exhibited lower HML-2 viral particle expression than those on a non-suppressive regimen [42, 44], indicating a direct or indirect link between HIV-1 activity and HML-2 expression. It is not clear if the reported decrease in HML-2 virion expression is due to HAART administration, the lack of HIV-1 replication, the reduced immune response to HIV-1 in those patients, or a combination of these.

There is limited information about which proviruses are expressed during HIV-1 infection. A recent investigation detected 15 type 1 and 18 type 2 HML-2 proviruses in

the plasma of HIV-1 infected patients in addition to apparent recombinant HML-2 viruses, whose presence would support the conclusion of ongoing HML-2 replication [46]. However, most recombinants in this study were detected using cloning of PCR amplicons, which was not capable of distinguishing recombinants arising during virus infection from artifactual recombination during PCR [46]. Potential recombinant HML-2 sequences were identified using single genome sequencing (SGS), a limiting dilution PCR and gold standard for characterizing viral genomes, in one HIV-1 patient who progressed to Hodgkin's lymphoma, suggestive of recombination during replication, but the central issue of HML-2 replication remains unresolved [46].

In addition to the investigation of HML-2 RNA in plasma, recent work has focused on examining HML-2 expression in HIV-1 infected and uninfected cells. HML-2 RNA and protein were expressed in cell culture models after HIV-1 infection or treatment with cell-activating agents like PMA/Ionomycin, the mitogen PHA and cytokine IL-2 [45, 86, 259]. The HML-2 LTR contains multiple binding sites for transcription factors associated with inflammation like NF-κB and NFAT, possibly explaining increased expression with cell activation [154]. HML-2 expression in the persistently HIV-1 infected T cell line KE37.1-IIIB was higher than its uninfected counterpart and appeared to predominantly originate from a single provirus present on chromosomal band 1q22 [259]. However, the use of activating agents CD3/CD28 and PMA/ionomycin on the Jurkat T cell line caused broad activation of multiple HML-2 proviruses and not just 1q22 [259]. Upregulation of HML-2 RNA has also been associated with the specific expression of the HIV-1 accessory proteins Tat and Vif in cell culture models, however the applicability of these findings to *in vivo* infection and their native concentrations is tenuous [48, 86, 112].

In primary cells from HIV-1 patients with advanced disease, HML-2 RNA was shown to be upregulated in CD4+ and CD8+ T cells, potentially indicating that an indirect effect of HIV-1 infection triggers HML-2 transcription as expression was not limited to only CD4+ T cells [179]. In this study, however, there was a weak inverse correlation between HML-2 RNA and an activated cell phenotype (CD38+HLA-DR+), contrary to the previous reports showing a positive impact of cell activation on HML-2 expression [179]. Similarly, in a study using primary cells, lymphocytes exposed to 5-azacytidine and PHA did not show an increase in HML-2 expression [58]. In CD3/CD28-activated uninfected lymphocytes, RNASeq profiling using PacBio sequencing of HML-2 *env* amplicons showed that >90% of HML-2 expression originated from three LTR Hs proviruses, namely 3q12.3, 1q22 and 1q23.3 [26]. However, whether expression changes in HIV-1 infected individuals was not investigated. The cause(s) of HML-2 expression during HIV-1 infection remain to be clarified along with the cell type(s) involved and specific loci expressed.

The production of HML-2 proteins during HIV-1 infection may lead to specific immune responses in some patients. Initial screens of HERV expression, including HERV-K (HML-2), HERV-H and HERV-F, found that individuals infected with HIV-1 generate higher antibody titers to HERV peptides than non-infected individuals [132, 144, 164, 229], though this has been contested [85]. An anti-HML-2 Env TM antibody isolated from an HIV-1 infected patient was shown to mediate killing of HIV-1 infected cells through an antibody dependent cell-mediated cytotoxcity (ADCC) mechanism, similar to a control anti-HIV antibody, though the efficiency of killing was dependent on cell donor and HIV-1 strain [165]. On a related note, anti-HML-2 Env TM antibodies

appeared more frequently in HIV-1 infected versus uninfected patients, however the prevalence of anti-HML-2 Env SU antibodies remained unchanged [164]. The reason for this observation remains to be clarified and the relationship between *in vitro* ADCC activity to *in vivo* control of HIV-1 infection is undefined.

In addition to antibody responses, cytotoxic CD8+ T cell responses to HERV epitopes were observed in HIV-1 infected patients but not uninfected controls [81]. In further analyses, there was an inverse correlation between the HIV-1 plasma viral load and anti-HERV CD8+ T cell responses in subjects, suggesting an association of HERV specific responses with controlled HIV-1 viral loads [81, 242]. The anti-HERV CD8+ T cells from HIV-infected individuals are capable of killing cells displaying HERV epitopes as well as HIV and SIV infected cells, though this was extremely variable [81, 112]. This finding indicates that if HERVs are expressed in HIV-1 infected cells they could mark the cells for elimination. HERV peptides could represent unchanging epitopes displayed on infected cells, presuming that HML-2 is not actively replicating and evolving in infected patients. Interestingly, some natural controllers, individuals who can control their HIV-1 levels without HAART, maintain an anti-HERV response in chronic HIV-1 infection when compared to control groups, including uninfected individuals and patients on HAART (both suppressive and non-suppressive) [222]. Accordingly, an HIV vaccine trial to induce immune responses to antigens from the HML-2 group was initiated [112].

Concerns about the efficacy of such a vaccine are relevant. The exact proviruses expressed during infection are not known and significantly, it is unclear if they are distinct from the proviruses expressed in different cell types within the body [72, 221]. In trials in rhesus macaques, an adenovirus-based vaccine could induce T cell responses to

HML-2 consensus Gag and Env and was found to be tolerable [204], however this immune response was not protective against SIV infection [224]. Lack of protection appeared to be due to the finding that SIV infection of rhesus macaque CD4+ T cells only lowly induced HML-2 expression [224], however another reason may be that the HML-2 proviruses reported to be expressed in humans are not present in macaques. The frequency of HIV-infected cells expressing HML-2 proteins is not yet established and the HML-2 proteins expressed could originate from proviruses not expressed in all individuals, making the success of such a vaccine dependent upon the frequency of the provirus in the human population. Importantly, HML-2 expression is seen in other diseases and physiological states that would preclude vaccine use.

The exact effects of HML-2 expression in HIV-1 pathogenesis remain to be elucidated but these initial results suggest an inhibitory role for anti-HML-2 immunity in HIV-1 pathogenesis, as well as a possible role for HIV-1 replication in the induction of HML-2 expression. The evidence provided thus far consists of reports of HML-2 RNA, protein and particle expression but does not ascribe a definitive role for HML-2 activity in contributing to or preventing HIV-1 disease progression in humans.

### *1.10 HML-2 interactions in the HIV-1 infected cell*

Based on *in vitro* studies, HML-2 expression during HIV-1 infection could affect HIV-1 replication if expressed in infected cells, though it is not clear if it would enhance or hinder HIV-1 pathogenesis. Integrase protein from the HML-2 provirus on chromosomal band 5q33.3 was shown to complement a defective HIV-1 integrase but virus infectivity was reduced to 3.7% of wild-type virus [256]. HML-2 protease

expressed from the same provirus appeared to be resistant to the effects of HIV-1 protease inhibitors *in vitro* and was able to cleave HIV-1 MA-CA peptides at the correct positions [251]. However, when tested in an *in vitro* system as a replacement for HIV-1 protease, it led to the production of non-infectious and improperly processed particles [181]. In another example of interaction between these distantly related viruses, HIV Rev [73], was reported to transport HML-2 genomic RNA out of the nucleus [274]. However, even if this were to occur *in vivo*, HML-2 RNA was not shown to be selectively packaged into HIV virions, likely due to HIV Gag not recognizing the packaging signal (ψ) on HML-2 RNA [280].

A recent study noted that a defective Env protein from the type 1 HML-2 provirus on chromosomal band 1q23.3 was able to pseudotype HIV virions *in vitro* [26]. If this activity occurs *in vivo*, it could potentially reduce the infectivity of pseudotyped HIV particles, as type 1 Env does not have an intact receptor-binding SU region. In contrast, the HML-2 Env encoded by the proviruses on chromosomal bands 6q14.1 and 7p22.1 were shown to counteract the effects of the enveloped virus restriction factor Tetherin, a protein which is incorporated into budding virions and tethers them to the cell surface [136, 173]. Though HIV-1 encodes an accessory protein named Vpu that counteracts Tetherin, expression of 6q14.1 and 7p22.1 Env could potentially allow for increased HIV-1 virus release in the presence of expressed HML-2 Env and negatively effect downstream innate immune signaling from Tetherin during HIV-1 infection [76]. In addition, the intact Env proteins from these same proviruses were shown to pseudotype SIV and HIV-1 virions, which could lead to changes in lentiviral tropism [59]. The

identities of the proviruses expressed during infection will be vitally important in determining their predicted effects on HIV-1 pathogenesis.

HML-2 Gag protein expressed from the sequence of a presumed HML-2 infectious progenitor was reported to co-assemble with HIV-1 Gag [166]. Co-assembly appeared to be dependent on the HML-2 NC domain, and reduced the release efficiency of HIV-1 virions [166]. A similarity between the extreme C-terminus of HML-2 CA and HIV-1 CA, a domain important for assembly of retroviral particles, has been noted previously [97]. However, *in vivo* evidence of co-assembly of retroviruses from these two different genera, namely HIV, a lentivirus, and HML-2, a betaretrovirus, has not been demonstrated. Thus, based on *in vitro* evidence, HML-2 expression could result in either enhanced or reduced levels of HIV-1 infection, likely dependent upon the complement of HML-2 proviruses expressed.

There remains a rare possibility that HML-2 activation leads to the formation of a replication competent virus. Recombination between the three proviruses located on chromosomal bands 6q14.1 (*gag-pro*), 8p23.1a (*pol*) and 7p22.1 (*env*) led to the production of a weakly infectious HML-2 virus *in vitro* [60]. Potentially, the virus would be able to integrate into target cells at transcriptionally active areas [25] and lead to cancer formation by established mechanisms, like insertional activation of a proto-oncogene [196]. HIV-infected patients show elevated rates of cancers, including Hodgkin's and non-Hodgkin's lymphomas, and cancers of the liver, lung and cervix [93]. There are HIV-associated malignancies that have no known cause, which include EBV-negative Hodgkin's disease and non-Hodgkin's B cell and T-cell lymphomas [24, 212]. Over half of AIDS-related systemic lymphomas are both EBV and HHV-8 negative and

the exact host parameters leading to genetic abnormalities and malignancy remain elusive [24]. Potentially, there could be a role for HML-2 expression in the pathogenesis of AIDS-associated cancers of unknown etiology.

An alternate method of HML-2 oncogenesis relates to the expression of retroviral proteins. Similar to HTLV-1 *tax*, if expressed, HML-2 *env*, *rec* and *np9* could contribute to cancer progression. Env expression has been implicated due to its assumed fusogenic and immunosuppressive activities [59, 169], while Rec and Np9 may play a role in initial events that lead to misregulation of cellular proliferation [3, 57]. If a suspected oncogenic HML-2 protein was encoded by an insertionally polymorphic provirus, its expression could relate to cancer progression in a subset of people and be determined through population studies. In this case and others, it is essential to determine the exact HML-2 proviral loci expressed in HIV-1 infection out of the many elements in our genome.

### *1.11 Potential mechanisms of HML-2 activation during HIV-1 infection*

While association of HML-2 RNA, protein and particle expression with HIV infection is increasing, the mechanism of proviral activation remains unclear. There are multiple possibilities to explain how HML-2 proviruses are activated during infection, mainly defined by their direct dependence on HIV infection of a cell or alternatively that activation is an indirect result of infection and may occur in uninfected cells.

Cell specific events could occur during HIV-1 infection that would allow for HML-2 expression. Potentially, HML-2 expression is directly reliant on HIV, where the proteins from HIV could promote transcription from HML-2 proviruses. The expression of HIV accessory proteins Tat and Vif were implicated in the upregulation of HML-2 RNA and

protein during *in vitro* HIV infection [48, 86, 112]. Neither Tat nor Vif has a defined function that would allow for a divergent virus like HML-2 to be transcribed. Tat allows for elongation of the nascent HIV transcript by binding a specific RNA structure on the HIV genome, while Vif inactivates the interferon-induced restriction factor and cytidine deaminase APOBEC3G [73, 160]. In terms of their association with HML-2 expression, Tat was proposed to recruit NF-κB to the HML-2 LTR, while no new function or explanation for Vif contribution was proposed beyond its role counteracting APOBEC3G [48, 86]. The overexpression of these proteins *in vitro* most likely does not relate to *in vivo* produced concentrations. Potentially, uncharacterized secondary functions exist for these proteins or other proteins encoded by HIV could promote HML-2 activation, though this remains to be seen.

Expression of HML-2 loci could be due to the position of HIV integration in the genome. Potentially, the promoter activities of the HIV LTR could increase transcription of HML-2 loci that are normally silenced by methylation and other epigenetic mechanisms [150]. HML-2 elements retained in our genome have been selected for many generations and are commonly located in transcriptionally silent regions, in opposite orientation to nearby genes [159], whereas HIV preferentially integrates into transcribed genes [218]. Therefore, this may be an unlikely mechanism due to lack of overlap in integration sites. Indeed, if expression of HIV-1 protein or presence of HIV integration is necessary, HML-2 transcription and protein expression would be limited to a subset of productively infected cells. Furthermore, the magnitude of HML-2 expression from these HIV-1 infected cells would be very high in order to detect such an event in mixed cell populations or in plasma.

Conversely, HML-2 expression may be due to an indirect mechanism and not directly dependent upon HIV infection of a cell. Chronic immune activation is commonly seen during HIV-1 infection, both on and off antiretroviral therapy [51]. Immune activation could lead to the presence of transcription factors able to bind the HML-2 LTR [154], or potentially cause epigenetic or methylation changes in host cells. This mechanism would also support the reports of HML-2 RNA and protein expression seen in multiple inflammatory disease states, as previously described. In HIV-1 infection, while most *in vitro* experiments show HML-2 upregulation after treatment with stimulating agents, *in vivo* correlations suggest that a high level of immune activation is correlated with a decrease in HML-2 RNA expression [45, 86, 179]. HML-2 expression in multiple cell types could be indicative of an indirect mechanism, or potentially a combination of direct and indirect causes.

The possible reactivation of HML-2 proviruses following immune stimulus is suggestive of both its potential use in extending gene networks reactive to specific signals (potentially to immune-related transcription factors) or the use of retroviral genes to block infection from circulating exogenous threats. However, it seems improbable that HML-2 expression is important on its own, particularly given the relatively recent introduction of some HML-2 proviruses to the human germ line. Rather, it is more likely that the transcriptional patterns observed reflect those of the ancestral exogenous HML-2 retrovirus. Since exact HML-2 virus cell tropism is not yet determined [59, 60, 134], HML-2 sensitivity to transcription factors relevant to immune activation could be indicative of its ancestral replication environment. In the case of the much older integrations, an understanding of the variety and ages of the elements responding to a

particular stimulus, their preserved transcriptional motifs and preserved retroviral genes will help clarify whether these elements are being expressed as a benefit to the host, as a byproduct of host gene regulation or actually reflect their history as infectious viruses.

### *1.12 Rationale for Study*

In this study, I proposed to identify the HML-2 proviruses expressed during HIV-1 infection, their distribution and level of expression, and the cell origin(s) of provirus activation in HIV-1 infected individuals. In PBMCs from HIV-1 infected and uninfected subjects, we explored whether HML-2 polymorphic loci were being expressed and compared overall patterns of HML-2 expression between populations, as differences in expression could indicate different disease outcomes based on inherited polymorphisms or provirus expression patterns. In addition, we investigated the source of provirus expression in PBMCs, which was essential to determine the probable mechanism of activation in HIV-1 infection. If provirus expression relied on an HIV-1 specific mechanism and were only expressed in infected cells, it would indicate that HML-2 expression is controlled by modulation of the host cell or is specific upregulated due to the presence of HIV-1, whereas the alternative result of HML-2 expression in uninfected cell types implies that HML-2 expression is a consequence of the host response to HIV-1 infection. Finally, this study established whether the expressed HML-2 sequences could be due to replication competent viruses, potentially arising by recombination among defective, but complementing proviruses, and critically assessed whether HIV-1 infection provides an environment for HML-2 virion production. Thus, the data presented here

provides greater detail about the phenomenon of HML-2 expression, its mechanism of activation and potential role as a factor in disease progression.

## Chapter 2: Materials and Methods

The methods in this chapter were published previously in:

Bhardwaj N, Maldarelli F, Mellors J, Coffin JM. "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*. 2014;88(19):11108-20.

Bhardwaj N, Montesion M, Roy F, Coffin JM. "Differential expression of HERV-K (HML-2) proviruses in cells and virions of the teratocarcinoma cell line Tera-1." *Viruses*. 2015;7(3):939-68.

### *2.1 Clinical Samples*

Plasma and peripheral blood mononuclear cell (PBMC) samples were obtained from Tufts Medical Center (TMC), the NIH Clinical Center and University of Pittsburgh (UPitt) under IRB approved protocols. All participants provided written informed consent. Patients were recruited from TMC in 2012-2014 and confirmed to be off antiretroviral medication, 18-65 years of age and free of confounding co-morbidities reported to have increased HML-2 expression, including cancer, schizophrenia, autoimmune disease or pregnancy. Blood samples from patients at TMC were drawn into BD Vacutainer Cell Preparation tubes with sodium citrate (BD, Cat# 362760) to allow for separation of plasma and PBMCs from the same blood draw. Samples were processed within 2 hours of blood draw according to manufacturer's instructions. Plasma was stored

at -80°C in 1ml aliquots and PBMCs were frozen in 5% dimethyl sulfoxide (DMSO) in fetal bovine serum at 5x106 cells/ml and stored in liquid nitrogen until analysis.

Participants at NIH were enrolled in clinical protocols (00-I-0110, 97-I-0082, 95-I-0072) approved by the NIAID Institutional Review Board (FWA00005897) administered at the NIH Clinical Center in Bethesda, Maryland. Participants at NIH provided written consent for participation, genetic analysis, sample storage and sharing with collaborators outside of intramural NIH. Two patients providing plasma samples were hepatitis C virus positive and one patient providing PBMCs was herpes simplex virus positive.

Blood samples from UPitt were drawn into BD Vacutainer K2EDTA tubes (BD, Cat# 366643) with EDTA. Plasma was separated from whole blood by centrifugation at 400xg followed by a second spin at 400xg, and was then stored at -80°C in 1.5ml aliquots. PBMCs were isolated from remaining (plasma-free) blood by Ficoll-Hypaque density gradient centrifugation. PBMC were frozen in 10% DMSO in fetal bovine serum at 5x106 cells/ml and stored in liquid nitrogen. Samples were shipped on dry ice and then transferred to liquid nitrogen storage until analysis. Three non-viremic patients providing PBMCs were hepatitis C virus positive.

All uninfected control samples used were drawn at the NIH Clinical Center or UPitt from adult volunteers aged 18-65 with no known co-morbidities. Control plasma and PBMC samples were stored as described above until thawed for analysis.

*2.2 HML-2 env primer design*

An HML-2 provirus alignment containing 91 sequences was downloaded from a previous publication [233] and analyzed in BioEdit Sequence Alignment Editor (Ibis

Biosciences, Carlsbad, CA). BioEdit was used to search for regions of high sequence conservation for development of an HML-2 env specific quantitative PCR (qPCR) and to determine which proviruses could be detected using the *env* specific qPCR based on sequence identity to primers.

*2.3 Cell culture*

The human teratocarcinoma cell line Tera-1 (ATCC, Cat# HTB-105) was grown at 37°C with 5% CO2 in McCoy's 5A media (Life Technologies, Cat# 16600-082), supplemented with 15% FBS (Atlanta Biologicals, Cat# S11195) and 1% Pen-Strep (Life Technologies, Cat# 15140-122). Passage-matched Tera-1 cells and Tera-1 supernatant were collected from 100mm cell culture plates. Culture supernatant was spun down for 5 min at 1200 x g and 0.22μm filtered to remove cellular debris. 1mL of 0.25% Trypsin-EDTA (Gibco, Cat# 25200-056) was added to the cell culture plate to remove cells and incubated at 37°C until detached. Cells were removed from the plate, washed 1X with 5mL Phosphate-buffered saline (PBS; Gibco, Cat# 14190), and pelleted for 5 min at 1200 x g. Dry pellets of Tera-1 cells and filtered supernatant samples were frozen at -80°C until the extraction procedures were performed.

*2.4 RNA extraction from plasma and cell supernatant*

RNA was extracted from HIV-infected patient plasma, control patient plasma and Tera-1 supernatant according to a modified version of a previously published protocol [161]. Plasma samples and filtered Tera-1 supernatants were thawed on ice the day of analysis. Plasma samples had a pre-spin at 2500xg performed at room temperature for 15

min to pellet any cellular debris. 200-500µl of plasma was diluted with an equal volume of Tris-buffered saline (TBS; Sigma, Cat# T5030) and 500uL of TBS was added to 1ml of Tera-1 supernatant in 1.7ml Eppendorf tubes. Virions were pelleted from plasma or supernatant at 21000xg for 1 hour at 4°C. For the Tera-1 virion RNA used for RNASeq analysis, pelleted virions from 3ml of Tera-1 supernatant were combined prior to downstream extraction.

The virion pellets were resuspended in 50µl of 5mM Tris-Hydrogen chloride (Tris-HCl; Invitrogen, Cat# 15568-025) and treated with 20mg/ml Proteinase K (Ambion, Cat#AM2548) for 30 min at 55°C. 200mL 6M guanidinium isothiocyanate (Sigma, Cat# 50983) and 10µl 20mg/ml glycogen (Roche, Cat# 10901393001) were added and vortexed, and the mixture was incubated for 5 min at RT. 280µl 100% isopropanol was added and the samples were centrifuged at 21000xg at 4°C for 35 min. The pellets were washed with 500µl 70% ethanol, centrifuged at 21000xg at 4°C for 15 min, and all ethanol removed through sequential spins and pipetting. The RNA pellets were air-dried for 2 min and resuspended in buffer (965µl 5mM Tris-HCl, 25µl RNasin or RNaseOUT, 10µl 0.1M dithiothreitol (DTT)). Tera-1 virion RNA was treated with 1.5U of DNase (Ambion, Turbo DNA-free kit, Cat# AM1907) for 45 minutes at 37°C and the enzyme was inactivated using the kit components. Plasma RNA was left untreated or treated with 1U of DNase for 30 minutes at 37°C and inactivated according to kit instructions.

For the HIV Roche Taqman v2.0 assay performed at UPitt, plasma HIV RNA was extracted using the automated COBAS AmpliPrep System v2.0 by Roche Molecular Diagnostics.

*2.5 Nucleic acid extraction from cells*

Unsorted patient PBMCs were removed from liquid nitrogen storage and heated in a 37°C water bath until almost thawed. Cells were resuspended in 10ml of Phosphate-buffered saline (PBS; Gibco, Cat# 14190) and centrifuged at 350xg for 5 min at 4°C. The cell pellet was resuspended in 2ml PBS and 10-40µl was used to obtain viable cell counts using trypan blue and a hemocytometer. 1-2 million cells were pelleted and used for extraction. Dry pellets of 5-10 million Tera-1 cells were removed from -80°C storage and thawed on ice. 1ml of TRIzol (Ambion, Cat# 15596-026) was added to the PBMC cell pellets and 5ml of TRIzol was added to the Tera-1 cell pellets for lysis for 5 min at RT. After lysis, the Tera-1 cell pellets were split into 5 samples for downstream extraction.

200µl chloroform was added to the TRIzol solution and mixed vigorously for 15 sec. Samples were incubated at RT for 3 min and then centrifuged at 12000xg for 15min at 4°C. 400-500µl of the aqueous phase were transferred into a new Eppendorf tube and an equal volume of 70% ethanol was added. RNA was extracted using a column purification kit (Ambion, PureLink RNA Mini, Cat# 1218301A) and treated according to manufacturer's instructions. RNA was eluted into 60µl of 5mM Tris-HCl (Invitrogen, Cat# 15568-025) and treated with 2U DNase for 1 hour at 37°C and the DNase inactivated according to manufacturer's instructions (Ambion, Turbo DNA-free kit, Cat# AM1907).

Sorted PBMCs were collected in PBS (Gibco, Cat# 14190) and kept on ice in a sorting tube (BD, Cat# 352063) until transferred out of the sorting facility. Cells were pipetted into a new Eppendorf tube, centrifuged and the RNA was extracted from the cell pellet according to manufacturer's instructions (Qiagen, AllPrep RNA/DNA Mini, Cat#

80204). The column was eluted with 40µl 5mM Tris-HCl (Invitrogen, Cat# 15568-025) and the eluate was passed over the column twice to concentrate RNA. RNA was treated with 2U DNase for 1 hour at 37°C and the DNase inactivated as above.

Nucleic acids (RNA and DNA) were extracted from PBMCs at UPitt following a previously described protocol [37], with sonication during the lysis step and again after resuspension of nucleic acids.

*2.6 Reverse transcription and quantitative PCR*

RNA was reverse transcribed using the HML-2 *env* reverse primer (sequence provided below) for the plasma HML-2 virion analysis and Tera-1 cell and virion analyses or with random hexamers (Invitrogen, Cat# N8080127) for the unsorted and sorted cell expression analyses to detect HML-2 and GAPDH transcripts. All assays included both RT positive (RT+) and RT negative (RT-) wells to detect the presence of contaminating DNA in the RNA sample and water RT+ and RT- controls were also run to detect reagent contamination. Reverse transcription reactions were set-up as recommended by the manufacturer (Invitrogen, Superscript III First Strand Synthesis, Cat# 18080-051) with the following cycling conditions: 50°C for 50 min, 85°C for 10 min, 4°C hold. cDNA was used for downstream qPCR reactions. Reverse transcription for the HIV qPCR analysis was performed as described previously [161].

RNA standards were prepared for the HML-2 qPCR and GAPDH by cloning the amplicon sequence into a vector with a T7 promoter (Invitrogen, pcDNA 3.1, Cat# K4900-01 or equivalent) and performing *in vitro* transcription (Ambion, MEGAscript T7, Cat# AM1334) to make copies of the qPCR amplicon. Standards for the HIV qPCR were

either supplied by Dr. Mary Kearney (NCI-Frederick) or prepared off a vector through *in vitro* transcription. IVT RNA was purified twice according to manufacturer's instructions (Ambion, MEGAclear, Cat# AM1908) and visualized on a denaturing gel to confirm the size of the IVT product. Standard RNA of known quantity was used in each assay to quantify the amount of RNA present in the test sample and was prepared in 10-fold serial dilutions and reverse transcribed at the same time as the test samples.

cDNA produced using an HML-2 gene specific primer was detected using an HML-2 qPCR specific to the TM region of the *env* gene. Specificity of this qPCR for HML-2 *env* was verified by melting curve analysis and sequencing of amplified products. The HML-2 primers used were: HML-2 *env* For (5' CTAACCATGTCCCAGTGATG 3') and HML-2 *env* Rev (5' GGAGACAGACTCATGAGCTTAGAA 3'). The primers were used at a final concentration of 300nM in a SYBR mastermix (Applied Biosystems, SYBR Green PCR MM, Cat# 4309155). HML-2 qPCR cycling conditions: 95°C 10 min; 95°C 15 sec, 57°C 15 sec, 72°C 45 sec, 74°C 15 sec, Plate read (x45 cycles); Melt Curve 55°C – 95°C. Each sample and water RT+ and RT- well was analyzed in triplicate using the HML-2 qPCR. The GAPDH qPCR was performed as described for the HML-2 assay. The GAPDH primers used are: GAPDH For (5' GTCAGTGGTGGACCTGACCT 3') and GAPDH Rev (5' TGCTGTAGCCAAATTCGTTG 3'). GAPDH cycling conditions: 95°C 10 min; 95°C 15 sec, 63.5°C 15 sec, 72°C 45 sec, 82°C 15 sec, Plate read (x45 cycles); Melt Curve 55°C – 95°C. HIV qPCR was performed as previously described [161].

To screen HIV-infected and uninfected PBMC RNA for DNA contamination prior to RNASeq analysis, samples were analyzed using a one-step quantitative PCR (Invitrogen,

Cat# 11746-100), which used the same HML-2 *env* primers and HML-2 *env* standards as described above. 2µl of DNase-treated RNA was used as input and run in triplicate or duplicate. The primers were used at a final concentration of 300nM (For) or 600nM (Rev) in the one-step mastermix. RT+ reactions were prepared using the supplied enzyme mix containing both SuperScript III reverse transcriptase and Platinum Taq DNA polymerase, while RT- reactions were prepared using only Platinum Taq DNA polymerase. One-step cycling conditions: 50°C for 3 min; 95°C 5 min; 95°C 15 sec, 60°C 30 sec, Plate read (x45 cycles); 40°C 1 min; Melt Curve 55°C – 95°C.

For the Roche HIV Taqman assay performed at UPitt, plasma HIV RNA was reverse transcribed and quantitated using the automated COBAS TaqMan System v2.0 by Roche Molecular Diagnostics, which has a quantitation limit of 20 copies/ml. Cellular HIV RNA and DNA were quantified using previously described protocols [36, 37]. For DNA quantification, following estimation of total nucleic acid concentration by NanoDrop 1000 (Thermo Scientific), samples were diluted to a final concentration of <170 ng/µL to prevent inhibition of qPCR. Eight replicates from each sample were assayed for total HIV-1 DNA or RNA using published qPCR methods with normalization for cellular input [36]. The 95% limit of detection (LOD) for the HIV-1 DNA or RNA was 5 copies per qPCR reaction.

*2.7 Quantitation of proviruses detected using HML-2 env primers*

Human genomic DNA (Applied Biosystems TaqMan Human Control DNA, Cat# 4312660) was supplied at 10ng/µl. 2-fold serial dilutions of human DNA were performed in 5mM Tris-HCl (Invitrogen, Cat# 15568-025). The standard assumption of 3pg DNA

per haploid genome was used to estimate the number of genomes present in a known quantity of DNA. 1.5µl of genomic DNA was loaded into the HML-2 qPCR in triplicate as described. Plasmid DNA standards (Invitrogen, pcDNA 3.1, Cat# K4900-01) containing the HML-2 *env* coding region from 7p22.1a (K108) were diluted to known quantities and used as standards to estimate number of HML-2 DNA copies in the different dilutions of human DNA. Three runs of this assay using the cycling conditions described for the qPCR were performed at different dilutions and the results used to estimate the number of HML-2 proviruses human DNA.

*2.8 Detection of K111 proviruses using HML-2 env primers*

The HML-2 *env* PCR was run using Platinum Taq HiFi reagent kit (Life Technologies, Cat#11304-011), HML-2 *env* primers at 300nM and 50ng DNA from human/rat hybrid cell lines containing only human chromosomes 14 (NA10479), 15 (NA11418), 16 (NA10567) and 19 (NA10449) obtained from the NIGMS Human/Rodent Somatic Cell Hybrid Mapping panel (Coriell). Chromosomes 14 and 15 contain K111 HML-2 proviruses with an *env* sequence, whereas chromosomes 16 and 19 contain non-K111 proviruses with an *env* sequence. HML-2 PCR cycling conditions: 95°C 2 min; 95°C 15 sec, 57°C 15 sec, 72°C 45 sec x 40; 4°C hold. PCR products were analyzed using 1.5% agarose gel electrophoresis and visualized with ethidium bromide and UV illumination.

*2.9 Flow cytometry*

PBMCs collected from 10 HIV-infected patients and 8 uninfected patients were

thawed in a 37°C water bath until icy and washed twice in buffer (PBS without $Mg^{2+}$ or $Ca^{2+}$, 2mM EDTA, 0.1% BSA). Cells were resuspended in buffer and stained with antibodies to CD3 (clone UCHT1), CD4 (clone RPA-T4), CD8 (clone RPA-T8), CD14 (clone MϕP9) and CD20 (clone 2H7). Antibodies and their isotype controls were procured from BD Biosciences (San Jose, CA) and BioLegend (San Diego, CA). Additionally, cells were stained with live/dead (L/D) stain (Molecular Probes, L/D Fixable Aqua Dead Cell Stain kit, Cat# L34957 or Propidium Iodide) to allow for sorting of live cells only, pre-incubated with an Fc receptor blocking solution to prevent non-specific binding of antibodies (BioLegend, Human TruStain FcX, Cat# 422302) and passed through a 40μM mesh strainer prior to sorting (BD, Cat# 352235) to minimize clumping. Samples were sorted on a BD Influx (405nm, 488nm, 635nm) using the BD Spigot software package. All cells were gated on absence of L/D stain and size (Pulse width x FSC, FSC x SSC) to exclude clumped cells and debris. For each patient tested, cells were sorted into lineages based on expression of the following markers: CD3+CD4+ (CD4+ T cells), CD3+CD8+ (CD8+ T cells), CD3-CD20+ (B cells), and CD3-CD14+ (Monocytes). Cells were sorted into PBS and kept on ice until RNA extraction. All flow cytometry analysis post-sort was performed using Summit version 4.3 build 2445 (Beckman Coulter, Fullerton, CA).

*2.10 RNASeq library preparation*

RNASeq libraries were prepared by the Genomics group in the Tufts University Core Facility. Illumina RNASeq libraries were prepared from Tera-1 cell RNA and human PBMC RNA using the TruSeq Stranded Total RNA kit with Ribo-Zero Gold (Illumina,

Cat# RS-122-2301), which removes ribosomal RNA from the test sample. ~1µg of sample RNA was depleted of rRNA and resulting RNA was incubated at 65°C for 5 min to avoid shearing (as recommended in alternate protocol), which should produce cDNA fragments ranging from 130-350 bases in length due to random priming. This step was followed by reverse transcription, end repair, an A-tailing reaction to add a single 3' A-overhang to the fragments and then ligation of barcoded sequencing adaptors with a T-overhang to bind these fragments. The library was amplified using adaptor-specific primers for 10-15 cycles of PCR.

An Illumina RNASeq library was prepared from the Tera-1 virion RNA using the NuGen Ovation v2 kit (NuGen, Part# 7102). This kit does not allow for strand-marking (dUTP incorporation) during cDNA synthesis. It takes low input samples like virion RNA, amplifies RNA and converts it to cDNA, which is then sheared using a targeted sonicator (Covaris M-Series, M220). Virion cDNA was sheared to 200-600bp, as determined using a BioAnalyzer. The Tera-1 virion library was prepared from the cDNA using end repair, A-tailing, barcoded adaptor ligation and library amplification as described above.

The Tera-1 cell and virion libraries were run together on the Illumina MiSeq benchtop sequencer (Tera-1 cell library = 95% input; Tera-1 virion library = 5% input) using the v3 kit that allows for paired-end (PE) reads up to 301 bases each in length (Illumina, Cat# MS-102-3001). PBMC libraries were multiplexed with 4 samples per run, using the same v3 kit. 26 million PE reads were generated for the Tera-1 cell library and 1.2 million PE reads for the Tera-1 virion library. On average, 5-7 million PE reads were generated for each PBMC sample.

*2.11 Alignment of RNASeq reads*

MiSeq reads from the PBMC, Tera-1 cell and Tera-1 virion libraries were trimmed to remove Illumina adaptor sequences, low quality reads (with a quality "Q" score <25) and reads shorter than 100 bases using the program Trimmomatic [21].

Trimmed paired-end reads from the libraries were aligned to the hg19 build of the human genome and to an HML-2 reference genome, which is a FASTA file containing the sequences of 93 2-LTR proviruses (4 are present only as solo LTRs in hg19, and 2 are present as pre-integration sites in hg19) and 943 solo LTRs for the Tera-1 alignments, or 96 2-LTR proviruses (4 are present only as solo LTRs in hg19, and 5 are present as pre-integration sites in hg19) and 975 solo LTRs for the PBMC alignments. Hg19 alignments for the Tera-1 cell and PBMC reads were performed using the –fr-firststrand option which allows for the strandedness of the read to be incorporated into the alignment data ("Plus stranded") or without this option ("Unstranded"). All Tera-1 virion alignments are unstranded since the library preparation did not allow for strand marking.

All alignments were performed using TopHat v2.0.10, which used Bowtie v2.1.0 as the underlying aligner [121, 129] and allowed for up to 2 mismatches to a mapping location for unique or multi-mapped reads. Output .bam files from the alignment were either (1) sorted and kept unfiltered ("Unfiltered") which retains reads that align to multiple targets as well as those that uniquely align to a single provirus or (2) sorted and filtered for uniquely aligned reads ("Unique Only") using SAMtools [137]. TopHat2 assigns uniquely aligned reads a mapping quality (MAPQ) score of 50 and these reads can be selected for from the total alignment using the SAMtools view –q 50 command.

The hg19 alignment indicates that aligned reads from the Tera-1 cell library had an average insert size of 180 bases (range: 98-522) as compared to 200 bases (range: 100-568) for the virion library and roughly 300 bases (range: 98->1000) for the PBMC libraries, which was determined using Picard Tools (Broad Institute) and/or QualiMap (Centro de Investigacion Principe Felipe, CIPF) for QC analysis of .bam files.

Alignments were visualized using the Integrative Genomics Viewer IGV v2.3.36 [249] and by using a custom track on the UCSC Genome Browser [120].

*2.12 Differential expression analysis on RNASeq reads*

For Tera-1 analysis, the Cufflinks module of Cufflinks v2.2.1 [194] was used to generate estimates of transcript abundance normalized to the length of the expressed gene, outputted as fragments per kilobase per million mapped reads (FPKM). Bam files representing the results of stranded, unstranded, filtered and unfiltered alignments were used as input. Hg19 transcript annotation files (GTF format) contained annotations for 87 HML-2 full proviruses, 4 proviruses present only as solo LTRs in hg19 and 947 solo LTRs in addition to cellular transcripts. HML-2 genome GTF files contained annotations for 93 included proviruses and 943 solo LTRs. Cufflinks was run using the standard default parameters or with the Multi-read correct –u parameter ("Multi-read Correct"), which assigns weighted FPKM values to loci with multi-mapping reads, based on an algorithm described previously [170].

FPKM values for individual HML-2 elements were used to calculate total HML-2 expression or packaging in the cells or virion by adding up the FPKM values from all HML-2 proviruses. From this number ("Total HML-2"), the percent abundance of each

HML-2 provirus compared to the total value was calculated as (provirus FPKM)/(total HML-2 provirus FPKM) x 100 to illustrate the relative contribution of individual proviruses to total HML-2 expression or packaging in the cell or virion.

For PBMC analysis, the CuffDiff and CuffNorm modules of Cufflinks v2.2.1 [194] were used to perform differential expression tests between the HIV-1 infected and uninfected populations and to export normalized FPKM values of relative gene expression, respectively. Bam files representing the results of stranded, unstranded, filtered and unfiltered alignments were used as input. Hg19 transcript annotation files used for expression analysis contained the annotations for HML-2 proviruses and solo LTRs as previously described, but either included or excluded cellular transcript annotations. HML-2 genome annotation files contained annotations for 96 HML-2 proviruses and 975 solo LTRs. As with Tera-1 RNASeq, total HML-2 expression was calculated by adding the relative expression values (FPKM) of individual proviruses together.

*2.13 MiSeq in-silico simulation*

Simulated MiSeq 250 base PE reads were generated from an HML-2 genome FASTA of 93 HML-2 proviruses to 20X coverage using the next-generation sequencing simulator program ART vVanillaIceCream-03-11-2014 [100]. Simulated reads were aligned back to the HML-2 genome using TopHat2 and were either kept “Unfiltered” or filtered for “Unique Only” alignment. FPKMs for each provirus were calculated using Cufflinks for both sets of FPKM values. As all proviruses were equally represented in the simulation, the average FPKM value for proviruses in the “Unfiltered” data set was used as the

comparator for the "Unique Only" data set in order to assess which proviruses were underrepresented after filtering.

*2.14 Phylogenetic analysis*

Neighbor-joining phylogenetic trees were created using MEGA6 [241]. Alignment of proviral or LTR sequence was performed using MUSCLE, an alignment tool native to the MEGA6 program. Neighbor-joining trees were constructed using the pairwise deletion option so that all available sites were used for comparison. The bootstrap values for the produced trees were the result of 1000 replicate tests. Distance was calculated using the p-distance method and the branch lengths correspond to the number of base differences per site.

*2.15 LTR Cloning and luciferase assays for 5' LTR activity*

The experiments described below were performed by Meagan Montesion, a graduate student in the Coffin lab, and supervised by Neeru Bhardwaj. The nucleotide sequences of selected HML-2 proviruses expressed in Tera-1 cells were obtained using the UCSC Table Browser [118]. Flanking primers to the 5' LTR of each provirus with were designed using Primer3 [255] and included restriction enzyme sites. These primers were used to PCR the 5' LTRs from Tera-1 genomic DNA using *Taq* DNA polymerase (Invitrogen, Cat# 10342-020). 5' LTRs were cloned in sense orientation into the pGL4.17[*luc2*/Neo] firefly luciferase vector (Promega, Cat. #E6721) and were screened for mutations prior to transfection.

For transfection, 24-well plates were seeded with Tera-1 cells at $1\times10^5$ cells/well in Opti-MEM reduced-serum media (Life Technologies, Cat# 31985-070). The pGL4-5' LTR vector was co-transfected with the pRL-SV40 internal control expressing *Renilla* luciferase (Promega, Cat# E2231) at a 30:1 ratio with Lipofectamine 2000 (Life Technologies, Cat# 11668-019), used according to the manufacturer's protocol. Background signal was assessed using non-transfected Tera-1 cells. 48 hours later, transfected cells were lysed and assayed using a dual-luciferase assay system (Promega, Cat. #E1910). The BioTek Synergy HT plate reader detected luminescence as relative light units (RLU) using Gen5 data analysis software (v2.03). Relative promoter activity for each 5' LTR was calculated after normalizing the firefly luciferase signal to the control *Renilla* luciferase signal.

*2.16 Statistical analysis and figure graphics*

All statistical analyses were performed using GraphPad Prism software version 6 (San Diego, CA) or Cufflinks software package v2.2.1. The exact statistical tests used are noted for each figure with p value $< 0.05$ to define statistical significance. qPCR data were analyzed using the specified non-parametric tests due to the small sample size and uncertainty about normal distribution in the population. Linear regression analyses using qPCR data were also performed in Prism, with p values noted where applicable. Parametric tests were performed on gene or provirus FPKM values generated from the PBMC RNASeq data sets as the expression values were fit to a negative-binomial distribution using CuffNorm. For analysis of "total HML-2" provirus expression, non-

parametric tests were performed since this value represents total expression from all HML-2 proviruses and is not expected to follow a normal distribution.

Graphics for scatter dot plots, pie charts and stacked bar graphs were generated using Prism. Figures 1-1, 1-2 and 5-1 were created using Microsoft PowerPoint. Age estimates used in Figure 4-2 and open reading frames for proviruses used in Figure 4-16 were obtained from a previous publication [233] or by inputting sequence into the NCBI ORF Finder [211]. Heatmaps were created using RStudio (RStudio: Integrated development environment for R, version 0.98.1060). FPKM values used for the heatmap were log-normalized using Decostand in RStudio Vegan and plotted using the RStudio Pheatmap module.

## Chapter 3: Characterization of HML-2 expression in HIV-infected individuals

The results in this chapter were published previously in:

Bhardwaj N, Maldarelli F, Mellors J, Coffin JM. "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. 2014;88(19):11108-20.

### *3.1 Lack of detection of HML-2 viral particles in plasma*

The HML-2 subgroup of HERV-K comprises >90 2-LTR proviruses per haploid human genome [233], many of which have accumulated inactivating mutations or deletions. To design a qPCR assay capable of detecting the majority of these proviruses, an alignment of all known HML-2 elements published by previous members of the lab was used and searched for regions of highest sequence conservation [233]. For primer design, *env* was targeted since this region is present on genomic transcripts as well as on singly spliced *env* transcripts. Also, in phylogenetic analysis, the sequence of HML-2 *env* has been shown to allow for distinct clustering of HML-2 proviruses from other related HML groups, thus targeting *env* should allow for specific PCR amplification of HML-2 proviruses [233].

A 119bp sequence located in the TM region of *env* was chosen as the target of our qPCR assay due to its high sequence conservation among both type 1 and type 2 HML-2 proviruses, which differ by the absence of a 292-bp stretch of sequence in the SU region of *env* [2, 143]. Based on primer sequence similarity to individual proviruses, an estimated 67 proviruses out of the 91 included in the alignment should be detected using

the *env* qPCR assay, including both recent and ancient integrations. The number of proviruses detected experimentally, 51 copies per haploid genome (Fig 3-1 A), was similar to, although slightly lower than, this prediction. Primer locations in other genomic regions would detect a similar number of proviruses, since many of HML-2 proviruses are missing substantial portions of their genome. The identity of the individual proviruses detected using the *env* qPCR assay was not investigated in detail, though the *env* primers were confirmed to detect K111 HML-2 proviruses, a subgroup of HML-2 proviruses known to be highly repeated due to post-integration duplication and which are potentially expressed at a high level during HIV-1 infection (Fig 3-1 B) [47]. RNA standards of known quantity, based on the K108 (7p22.1a) *env* sequence, were created to quantitate the copy number of HML-2 *env* RNA signal from cell and plasma specimens. Assay performance between multiple runs of the standards showed high repeatability, high PCR efficiency and a linear dynamic range of $10^7$ down to 10 copies (Fig 3-1 C).

Prior reports [42-44, 46, 47] presented data in support of the idea that HML-2 virions were released into the blood of HIV-1 infected patients. To see if we could confirm this result, plasma was collected from two different clinical centers and tested for the presence of HML-2 virions. De-identified plasma from HIV-1 infected patients (n=9) was collected from the archives at the NIH Clinical Center. Additional patients (n=6) were recruited in a clinical study ongoing at Tufts Medical Center specifically for HIV-1 infected patients naïve to therapy or off ART for >2 months. All samples tested were from patients off ART at the time of collection to exclude any possible effect of ART on HML-2 detection [42, 44]. Sample characteristics are detailed in Table 3-1.

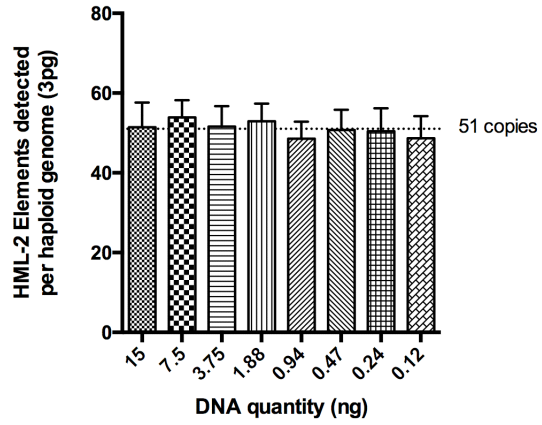
RNA was extracted from pelleted plasma and treated with recombinant DNase I to degrade any contaminating genomic DNA that might have led to a false HML-2 RNA signal in the *env* qPCR assay. DNase-treated RNA was reverse transcribed and analyzed for virus expression using the HML-2 *env* qPCR or a modified HIV single copy assay (SCA) [161]. A previously published report estimates that ~95% of patients infected with HIV-1 have detectable HML-2 RNA in their plasma [43]; and, moreover, that extremely high levels of HML-2 RNA are present in the plasma of those patients, ranging from $10^3$ to $10^{10}$ copies per ml [42, 44, 46, 47]. Contrary to these reports, we could not detect HML-2 RNA in the plasma of any HIV-1 infected patient tested (Fig 3-2 A). It is possible that the DNase used contained contaminating RNases, or that the DNase was not properly inactivated, leading to a reduction in HML-2 detection. Therefore, a comparison between DNase-treated and untreated RNA was performed with and without reverse transcriptase to determine if an HML-2 RNA signal could be measured in the absence of DNase. However, even without DNase treatment, an HML-2 RNA signal was not detected and the RT+ and RT- wells yielded equivalent copy values, suggestive of only DNA template being present (Fig 3-2 A). In comparison, the use of DNase did not have a significant effect on the detection of HIV RNA (Fig 3-2 C).

The cellular gene GAPDH could be detected when the untreated "RNA" was used at a qPCR template but not when DNase-treated RNA was used, consistent with the presence of a contaminating DNA template (Fig 3-2B). However, the strength of the signal is much weaker for GAPDH as compared to the HML-2. In Fig 3-2 A, the HML-2 signal in estimated copies/mL was determined using an RNA standard. Based on control experiments, the use of a DNA standard would show a 10-fold reduction in the estimated

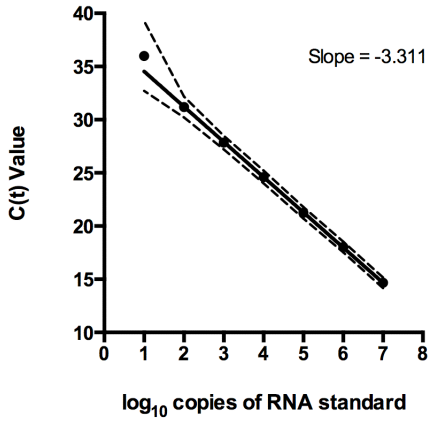
**Figure 3-1. qPCR detection of HML-2 proviruses.**

(A) The number of HML-2 proviruses detected per haploid genome using *env* qPCR was quantified using dilutions of human genomic DNA at known concentration (using the estimate of 6pg DNA/cell and 3pg DNA/haploid genome) and determining the copy number of HML-2 DNA at these different dilutions using plasmid DNA standards containing the 7p22.1a *env* sequence. Mean and standard deviation are plotted for each dilution, where the mean is the average of replicate wells from 3 assays. (B) Detection of K111 HML-2 proviruses was determined by running the HML-2 *env* PCR on human/rat hybrid DNA containing only chromosomes 14, 15, 16 or 19. The K111 provirus named K105 has sequence that is a perfect match for HML-2 *env* PCR primers. Chromosomes 14 and 15 contain K111 HML-2 proviruses with *env* sequence, as well as non-K111 HML-2 proviruses that lack amplifiable *env* sequence due to deletions. Chromosomes 16 and 19 contain only non-K111 HML-2 proviruses with amplifiable *env* sequence. (C) RNA standard serial dilutions were reverse transcribed in duplicate and the cDNA was assayed using the SYBR green *env* qPCR. Mean cycle threshold values (Ct, y-axis) and 95% CI are plotted for each dilution of the RNA standard (x-axis). Ct values are compiled from multiple assays, where n=41 for 10-$10^6$ copies and n=26 for $10^7$ copies. The slope of the line is -3.311, correlating to a PCR efficiency of 100.46%.

A


HML-2 Elements detected per haploid genome (3pg)
80
60
40
20
0
51 copies
15
7.5
3.75
1.88
0.94
0.47
0.24
0.12
DNA quantity (ng)

C


C(t) Value
40
35
30
25
20
15
10
Slope = -3.311
0
1
2
3
4
5
6
7
8
log10 copies of RNA standard

B

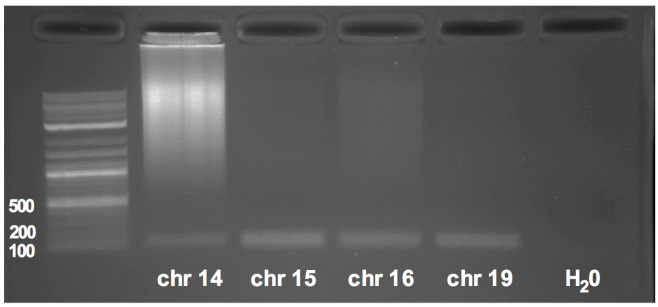

500
200
100
chr 14
chr 15
chr 16
chr 19
H20

**Table 3-1. HIV-1 infected patient characteristics for plasma and PBMCs analyzed in Figures 3-2, 3-3, 3-4 and 3-6.**

| NIH Clinical Center and TMC* | Plasma = 15 | | PBMC = 19 | |
|---|---|---|---|---|
| | n | % | n | % |
| *Demographic Characteristics* | | | | |
| Male | 14 | 93 | 18 | 95 |
| Caucasian | 6 | 40 | 9 | 47 |
| African-American | 4 | 27 | 3 | 16 |
| Hispanic | 5 | 33 | 6 | 32 |
| Asian Pacific-Islander | 0 | 0 | 1 | 5 |
| | Median | Range | Median | Range |
| Age (years) | 42 | 25.1 – 48.0 | 35.9 | 19.5 – 51.9 |
| *HIV Characteristics* | | | | |
| CD4 Count | 523 | 220 – 1105 | 661 | 210 – 1105 |
| %CD4 Cells | 24.8 | 10.7 – 55.7 | 29 | 9 – 55.7 |
| RNA Viral Load ($log_{10}$ copies/mL) | 4.2 | 1.95 – 5.6 | 4.11 | 1.95 – 5.55 |

Note: One viremic HIV-infected patient was off therapy during blood collection. All other patients are treatment-naïve and viremic.
*TMC, Tufts Medical Center.

copy number since the use of an RNA standard overestimates the abundance of DNA due to the <100% efficiency of RT. In addition, our GAPDH primers detect 0.91 copies of GAPDH per haploid genome, as compared to 51 copies of HML-2 per haploid genome (which includes K111 and non-K111 HML-2 proviruses). Due to this difference in assay sensitivity, we expect to see a 56-fold difference in signal between the HML-2 and GAPDH qPCR. Thus, in total, we expect up to a 66-fold difference in estimated GAPDH DNA/mL as compared to HML-2, explaining the small signal for GAPDH in comparison to HML-2, but supporting the conclusion that the signal is derived from a DNA template.

Our failure to detect HML-2 RNA was not due to improper extraction from the plasma samples, since HIV RNA was detected in the clinical specimens using the same RNA preparations that were used for the HML-2 assay (Fig 3-2 C). Additionally, HIV was detected with or without DNase treatment of viral RNA, although there was a ~4x reduction in RNA levels after DNase treatment (Fig 3-2 C). In addition, the HML-2 RNA extraction and detection procedures were validated by our successful detection of HML-2 RNA in supernatants from the teratocarcinoma cell line Tera-1, which is known to produce HML-2 virions [16, 27, 142, 202]. In these experiments, differing amounts of HML-2 virions were spiked into 300µl of HML-2 negative human plasma until the HML-2 RNA signal was indistinguishable with qPCR background. Using this approach, the functional limit of detection with our extraction and qPCR methods occurred when 50-80 copies of HML-2 RNA were present in the initial sample.

In case the extraction and detection of HML-2 virions failed for unknown reasons, previously published extraction methods, which involved DNase treatment of plasma followed by RNA extraction [46] or use of the commercially available Qiagen Viral RNA

kit [42, 46], followed by detection using a published qPCR *env* primer set [46] were performed on a subset of clinical samples. Similar to Fig 3-2 A, these alternative methods did not lead to the detection of HML-2 RNA and showed equivalent signal in RT+ and RT- wells. Thus, in the group of patients tested, we could find no evidence for HML-2 virions in plasma.

By contrast to the lack of HML-2 RNA detection in plasma, our experiments did show a significant difference in the levels of HML-2 DNA in the plasma of HIV-1 infected patients as compared to controls (Mann-Whitney, *p=0.02, Fig 3-2 A). Based on the estimated copies of HML-2 proviral DNA detected per haploid genome (Fig 3-1 A) and the difference in signal intensity from HML-2 RNA versus DNA in the *env* qPCR assay, the control plasma DNA level was calculated to be equivalent to ~11 lysed cells per ml, whereas the DNA level from HIV-1 infected patients was about 2-fold higher with an estimated ~23 lysed cells per ml of plasma (Fig 3-3 A). The source of the cell DNA is unknown, but it could be related to immune surveillance and/or infected cell killing during active HIV replication in the absence of ART. However, the level of HML-2 DNA signal in the plasma was not strongly associated with HIV RNA levels, CD4+ T cells/μl or % CD4 T cells in the HIV-infected patients (Fig 3-3 B-D).

*3.2 Upregulation of HML-2 transcription in peripheral blood mononuclear cells*

Although HML-2 virions were not detected in the plasma of the tested HIV-1 infected or uninfected patients, it was possible that HML-2 was actively transcribed in PBMCs [179]. As with the plasma samples, PBMCs were obtained from archived samples from the NIH Clinical Center (n=13) and also from an ongoing clinical study at Tufts Medical

Center (n=6), where all patients were verified to be off ART. Sample characteristics are listed in Table 3-1. PBMCs were assessed for viability and 1-2 million cells were used for total RNA extraction, followed by DNase treatment, reverse transcription and analysis for total copy number using qPCR assays for *env* and the reference GAPDH transcripts.

PBMCs isolated from HIV-infected patients showed a significant upregulation in HML-2 transcription relative to GAPDH, as compared to uninfected controls (Mann-Whitney, ***$p<0.0001$, Fig 3-4 A). The extent of HML-2 transcription was not significantly associated with any of the tested HIV-1 disease markers, including plasma HIV RNA levels, CD4+ T cells/μl or % CD4 T cells (Fig 3-4 B-D).

**Figure 3-2. Absence of HML-2 virions in plasma from HIV-1 infected patients.**

(A) Plasma from viremic HIV-1 infected and uninfected patients was centrifuged at 21000xg to pellet virions. RNA extraction was followed by DNase treatment (+ DNase) or no treatment (No DNase). RNA was reverse transcribed (RT+) along with no RT controls (RT-). RT+ and RT- wells were analyzed in triplicate using the HML-2 *env* qPCR. The dotted line represents the limit of detection (730 copies), and the geometric mean is plotted for each group of samples (*p=0.02, Mann-Whitney). (B) RNA from viremic HIV-1 infected and uninfected patients was reverse transcribed and copies of the cellular gene GAPDH were detected in the cDNA using a GAPDH specific SYBR qPCR. DNA standards to quantify GAPDH DNA were prepared from the GAPDH amplicon cloned into a standard cloning vector. Each point represents the average of triplicate wells from one individual. (C) The same RNA samples used for HML-2 detection were analyzed using a modified HIV single copy qPCR. The dotted line represents the limit of detection (200 copies) and the geometric mean is plotted for both HIV-1 infected patient groups (p=0.06, Wilcoxon signed rank test).

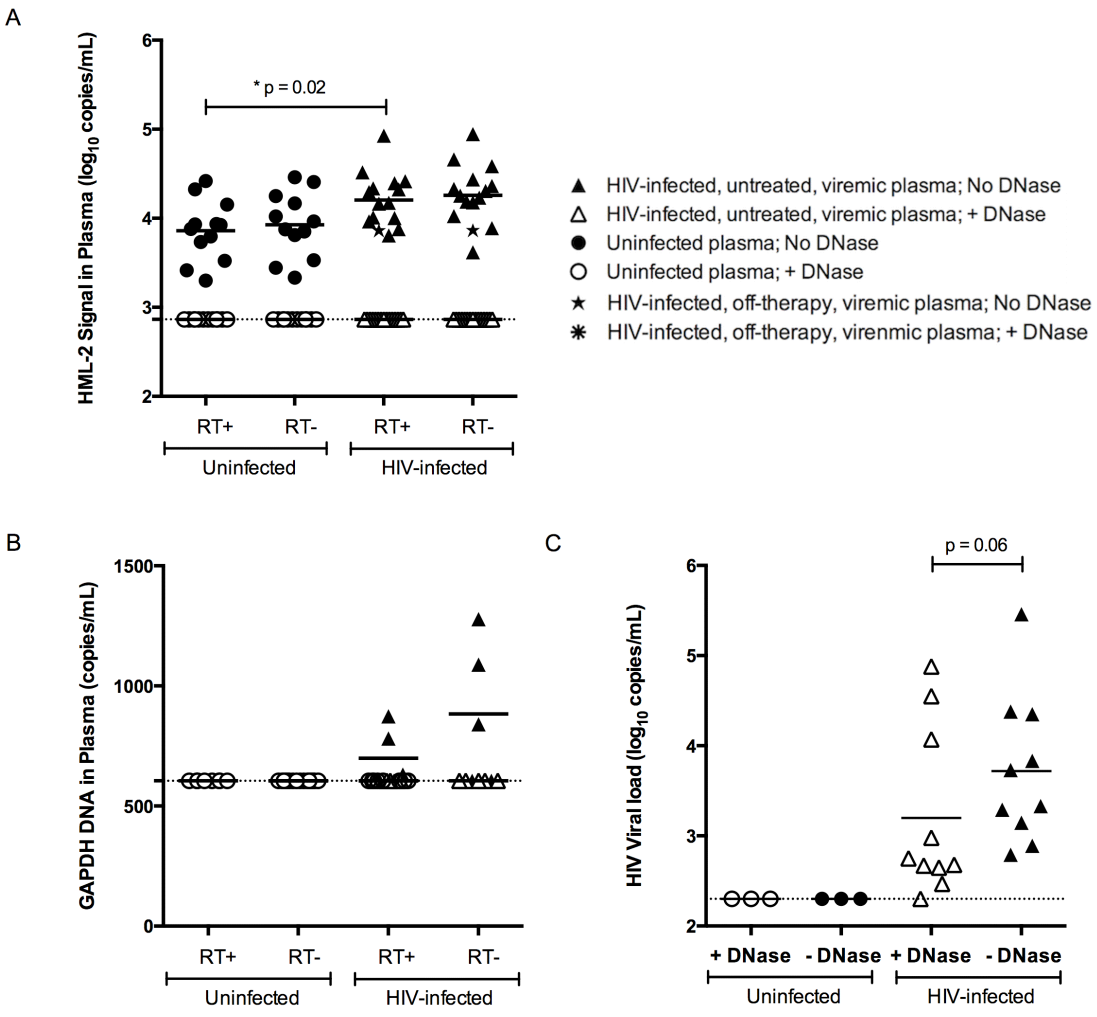

A
HML-2 Signal in Plasma (log10 copies/mL)
* p = 0.02
RT+
RT-
RT+
RT-
Uninfected
HIV-infected
HIV-infected, untreated, viremic plasma; No DNase
HIV-infected, untreated, viremic plasma; + DNase
Uninfected plasma; No DNase
Uninfected plasma; + DNase
HIV-infected, off-therapy, viremic plasma; No DNase
HIV-infected, off-therapy, virenmic plasma; + DNase
B
GAPDH DNA in Plasma (copies/mL)
C
HIV Viral load (log10 copies/mL)
p = 0.06
+ DNase
- DNase
+ DNase
- DNase
Uninfected
HIV-infected

**Figure 3-3. Presence of DNA in the plasma of HIV-1 infected patients.**

(A) The amount of HML-2 DNA in plasma from healthy controls and HIV-1 infected subjects was estimated for the RT- wells shown in Fig 3-2 A. A conversion factor between RNA and DNA signal in the *env* qPCR was determined by running RNA and DNA standards simultaneously. Cell number was calculated from the resulting DNA copy number (102 copies HML-2 DNA/cell). Means are plotted for each group. (B-D) The level of HML-2 plasma signal is plotted against the level of HIV-1 viremia (B), the CD4+ T-cell count (C), and the % CD4+ T-cells (D). The p values for linear regression are shown in the lower right hand corner of B-D. Refer to the figure legend shown for Fig 3-2.

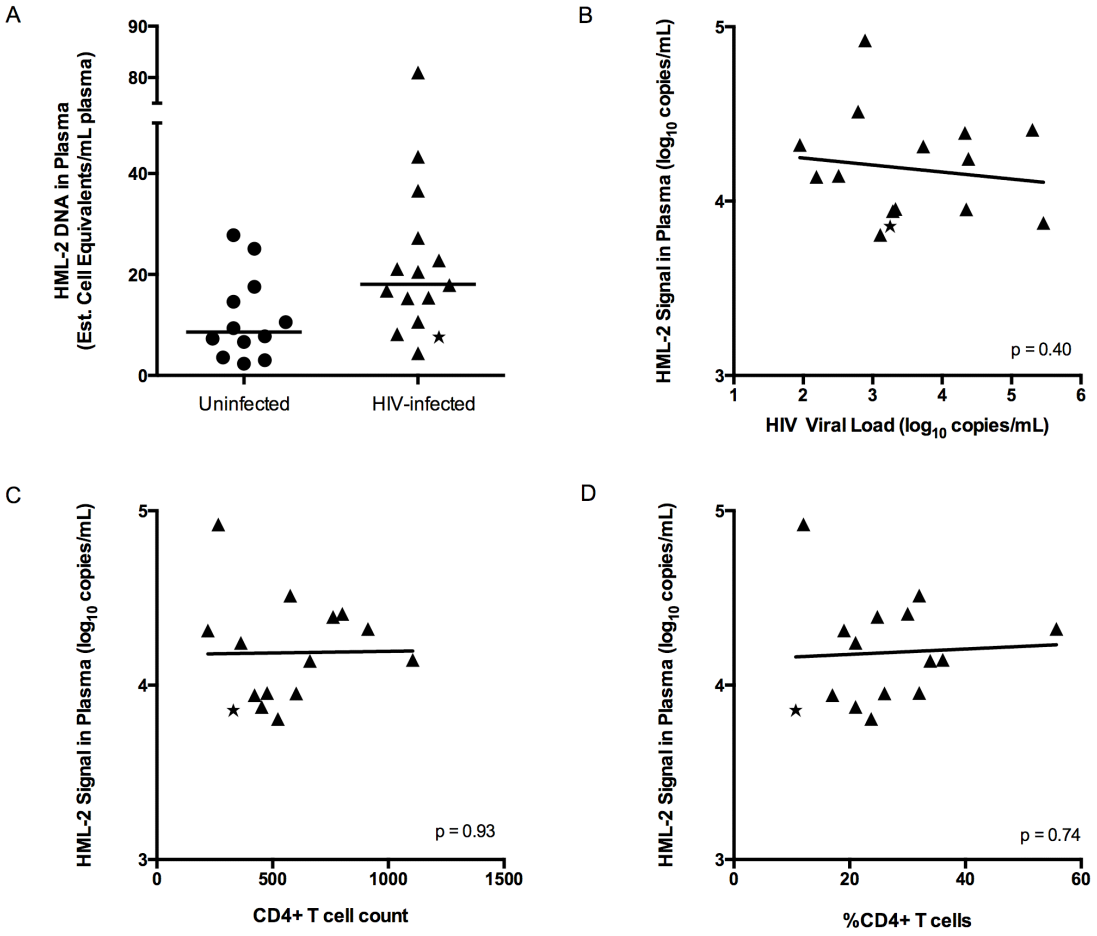
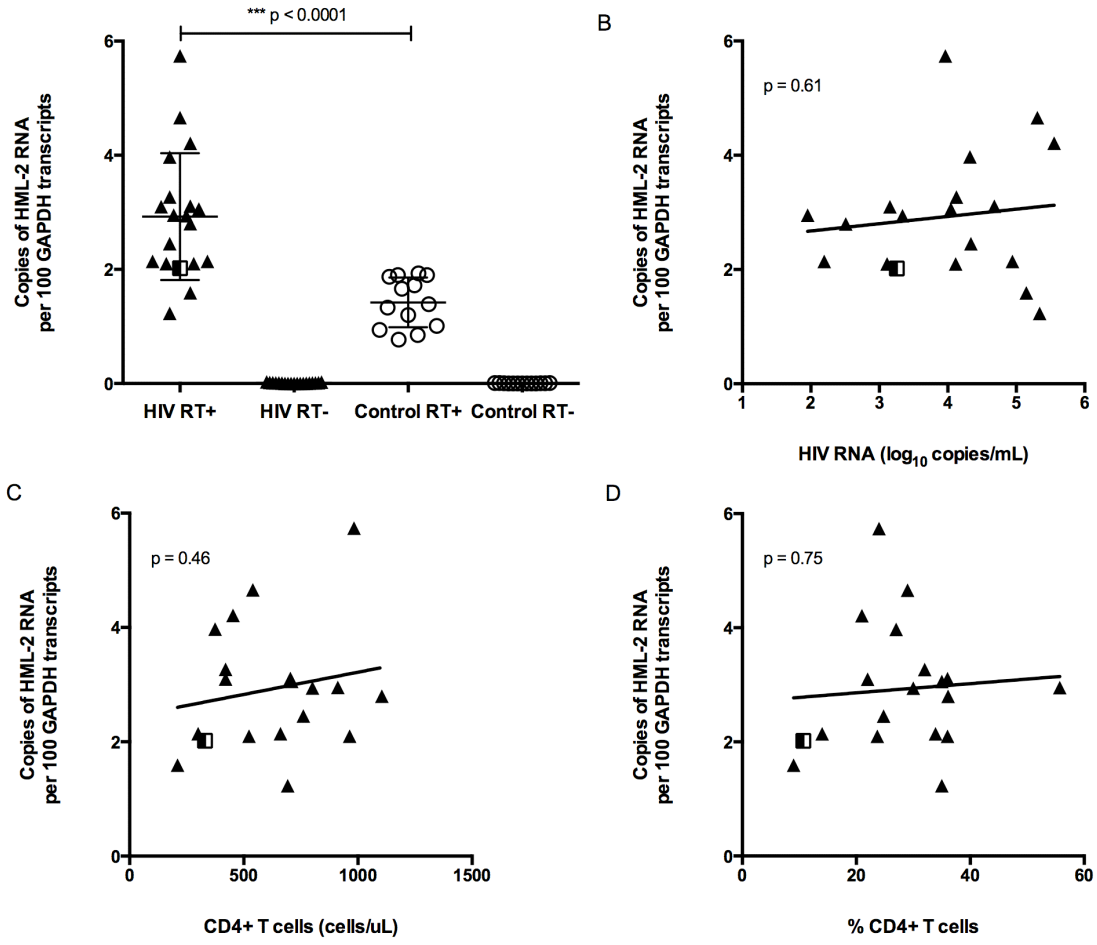

A
HML-2 DNA in Plasma
(Est. Cell Equivalents/mL plasma)
90
80
40
20
0
Uninfected
HIV-infected
B
HML-2 Signal in Plasma (log10 copies/mL)
5
4
3
1
2
3
4
5
6
p = 0.40
HIV Viral Load (log10 copies/mL)
C
HML-2 Signal in Plasma (log10 copies/mL)
5
4
3
0
500
1000
1500
p = 0.93
CD4+ T cell count
D
HML-2 Signal in Plasma (log10 copies/mL)
5
4
3
0
20
40
60
p = 0.74
%CD4+ T cells

**Figure 3-4. HML-2 RNA expression in PBMCs from HIV-1 infected patients.**

(A) HML-2 RNA was quantitated in PBMCs collected from viremic HIV-1 infected subjects as well as uninfected controls. Relative expression of HML-2 was assessed for each sample in triplicate by comparing HML-2 copy number to the reference gene GAPDH copy number. RT- controls are shown for all samples. The data are reported as copies of HML-2 RNA per 100 copies of GAPDH RNA (***$p<0.0001$, Mann-Whitney). Mean and standard deviation are plotted for each group. (B-D) The extent of HML-2 upregulation in PBMCs is plotted against (B) HIV plasma RNA, (C) CD4 T cells/uL, or (D) %CD4 T cells. The p value for linear regression (B-D) is shown in the upper left hand corner.


*** p < 0.0001
Copies of HML-2 RNA per 100 GAPDH transcripts
HIV RT+
HIV RT-
Control RT+
Control RT-
B
p = 0.61
HIV RNA (log10 copies/mL)
C
p = 0.46
CD4+ T cells (cells/uL)
D
p = 0.75
% CD4+ T cells
HIV-infected, untreated, viremic patient PBMC
HIV-infected, off-therapy, viremic patient PBMC
Uninfected control PBMC

*3.3 Relationship of HML-2 expression to HIV-1 replication*

To more directly test the correlation of HML-2 expression and HIV replication and determine if HIV replication is necessary for the apparent upregulation of HML-2 transcription, a blinded study was performed to assess HML-2 RNA levels in PBMCs from 15 HIV-1 patients on therapy, 10 without viremia (<75 copies of RNA/ml), and 5 with detectable viremia, as compared to uninfected controls (n=4) (Table 3-2). Of the 5 patients with viremia, 3 were off ART and 2 were on failing ART regimens (Table 3-3). Compared to the uninfected controls, there was upregulation in HML-2 RNA in both groups of patients on ART, but only to a significant level in the aviremic (Mann-Whitney, *p=0.02) and not the viremic patients (Mann-Whitney, p=0.29) (Fig 3-5 A). The lack of statistical significance for viremic group could have been due to the smaller sample size (n=5). When considering only the patients on therapy, whether viremic or aviremic, the level of HML-2 upregulation seen in the treated HIV-1 infected patients was similar in magnitude to untreated HIV-1 infected patients (Fig 3-4 A, Fig 3-5 B). Thus, the use of antiretrovirals and level of HIV replication did not appear to have a major effect on HML-2 RNA transcription (Fig 3-5 C-E). In addition, there was no association of HML-2 RNA with HIV DNA levels (Fig 3-5 F).

*3.4 Detection of HML-2 RNA in sorted PBMCs*

We next investigated the cell source of the HML-2 RNA expression to begin to elucidate the mechanism(s) governing HML-2 provirus activation in HIV infected patients. Attempts to isolate HML-2 expressing cells using a commercially available antibody to HERV-K Env were unsuccessful; therefore, PBMCs were sorted by cell type

and then assayed for HML-2 RNA. Specifically, PBMCs isolated from patients described in Table 1 were subjected to live cell sorting into the following PBMC cell subsets: CD4+ T cells (CD3+CD4+), CD8+ T cells (CD3+CD8+), B cells (CD3-CD20+) and monocytes (CD3-CD14+). HML-2 RNA levels were determined as described for total PBMCs.

Consistent with the lack of correlation of HML-2 RNA levels with HIV replication, sorted CD4+ T cells were not enriched for HML-2 RNA as compared to other cell subsets (Fig 3-6 A). Interestingly, no cell type was significantly enriched for HML-2 RNA transcription; in fact, all PBMC subsets tested showed detectable HML-2 expression in both the HIV-1 infected and uninfected populations, although to different extents (Fig 3-6 A-D). This result is consistent with previous assessments showing HML-2 expression in blood cells [220, 221]. In all cell types, the HIV infected patients exhibited a slightly greater level of HML-2 expression, with the greatest difference in monocytes (Mann-Whitney, $p=0.18$, Fig 3-6 C). However, no significant correlation was found between the percent live monocytes found in PBMCs compared to HML-2 RNA upregulation in total PBMCs ($p=0.56$, Fig 3-6 G). Indeed, in no cell type did the difference reach statistical significance, even though there was a significant difference when unsorted PBMCs were analyzed (Fig 3-4 A).

**Table 3-2. HIV-1 infected patient characteristics for PBMCs analyzed in Figure 3-5.**

| UPitt* | Non-Viremic = 10 | | Viremic = 5 | |
|---|---|---|---|---|
| | n | % | n | % |
| *Demographic Characteristics* | | | | |
| Male | 7 | 70 | 5 | 100 |
| Caucasian | 3 | 30 | 0 | 0 |
| African-American | 7 | 70 | 4 | 80 |
| Hispanic | 0 | 0 | 1 | 20 |
| Asian Pacific-Islander | 0 | 0 | 0 | 0 |
| | Median | Range | Median | Range |
| Age (years) | 51.5 | 25 – 58 | 52 | 29 – 53 |
| *HIV Characteristics* | | | | |
| CD4 Count | 617 | 416 – 1373 | 874 | 244 – 1091 |
| %CD4 Cells | 31.1 | 20.1 – 52.1 | 31.9 | 9.3 – 47.4 |
| iSCA** Viral Load ($log_{10}$ copies/mL) | ND | <-0.2 – 1.4 | 3.44 | 2.75 – 5.32 |

Note: All non-viremic patients were on therapy. Viremic patients were either treatment-naïve, off-therapy, or on therapy at the time of blood collection. Refer to Table 3-3 for treatment details.
* University of Pittsburgh Medical Center
** iSCA is a modified single copy assay (SCA), which detects a small region in the integrase portion of HIV-1 RNA.
ND = not determined. Four patients were below the limit of detection for the iSCA assay.

**Table 3-3. Therapy regimens for patients analyzed in Figure 3-5.**

| Patient | Sex | Ethnicity | Status | Treatment | Co-morbidities | HIV Plasma RNA |
|---|---|---|---|---|---|---|
| 1 | Male | Caucasian | Non-Viremic | Raltegravir, Abacavir/Lamivudine | none | < 0.6 |
| 2 | Female | Caucasian | Non-Viremic | Ritonavir, Emtricitabine/Tenofovir, Darunavir | none | < 0.6 |
| 3 | Female | African American | Non-Viremic | Efavirenz/Emtricitabine/ Tenofovir | Hepatitis C | < 0.7 |
| 4 | Male | African American | Non-Viremic | Efavirenz/Emtricitabine/ Tenofovir | none | < 0.6 |
| 5 | Male | African American | Non-Viremic | Atazanavir, Ritonavir, Emtricitabine/Tenofovir | none | 3.8 |
| 6 | Female | African American | Non-Viremic | Atazanavir, Ritonavir, Emtricitabine/Tenofovir | none | 0.6 |
| 7 | Male | African American | Non-Viremic | Efavirenz, Emtricitabine/Tenofovir | Hepatitis C | 2.5 |
| 8 | Male | African American | Non-Viremic | Atazanavir, Ritonavir, Emtricitabine/Tenofovir | Hepatitis C | 1.9 |
| 9 | Male | African American | Non-Viremic | Emtricitabine, Tenofovir, Raltegravir | none | 25.7 |
| 10 | Male | Caucasian | Non-Viremic | Zidovudine, Emtricitabine/Tenofovir, Etravirine, Raltegravir | none | 4.3 |
| 11 | Male | African American | Viremic | Ritonavir, Emtricitabine/Tenofovir, Darunavir | none | 866.7 |
| 12 | Male | African American | Viremic | Treatment naïve | none | 17,341 |
| 13 | Male | Latino | Viremic | Elvitegravir/Cobicistat/ Emtricitabine/Tenofovir | none | 2,774 |
| 14 | Male | African American | Viremic | Treatment naïve | none | 206,800 |
| 15 | Male | African American | Viremic | Not on medication | none | 564 |

**Figure 3-5. HML-2 RNA in PBMCs from HIV-1 infected patients on antiretroviral therapy.**

(A) HML-2 RNA was quantitated in PBMCs collected from HIV-1 infected subjects on antiretroviral therapy without detectable viremia (HIV RNA <75 copies) on therapy (open squares), from patients with viremia who were either on therapy (filled squares), currently off therapy (half-filled squares), or naïve to therapy (filled triangles) and from uninfected controls (open circles) as described in the legend. RT- wells were below the limit of detection (data not shown). The data are reported as copies of HML-2 RNA per 100 copies of GAPDH RNA. (*p=0.02, Mann-Whitney). Mean and standard deviation are plotted for each group. (B) Data from patients on therapy with or without associated viremia are combined into one data set to show the effect of antiretroviral therapy on HML-2 upregulation (C-F) The level of HML-2 RNA expression in PBMCs is plotted against (C) HIV plasma RNA measured by TaqMan assay, (D) HIV plasma RNA measured by iSCA assay, (E) cellular HIV RNA, or (F) cellular HIV DNA. The p value for linear regression is shown in the upper left hand corner.

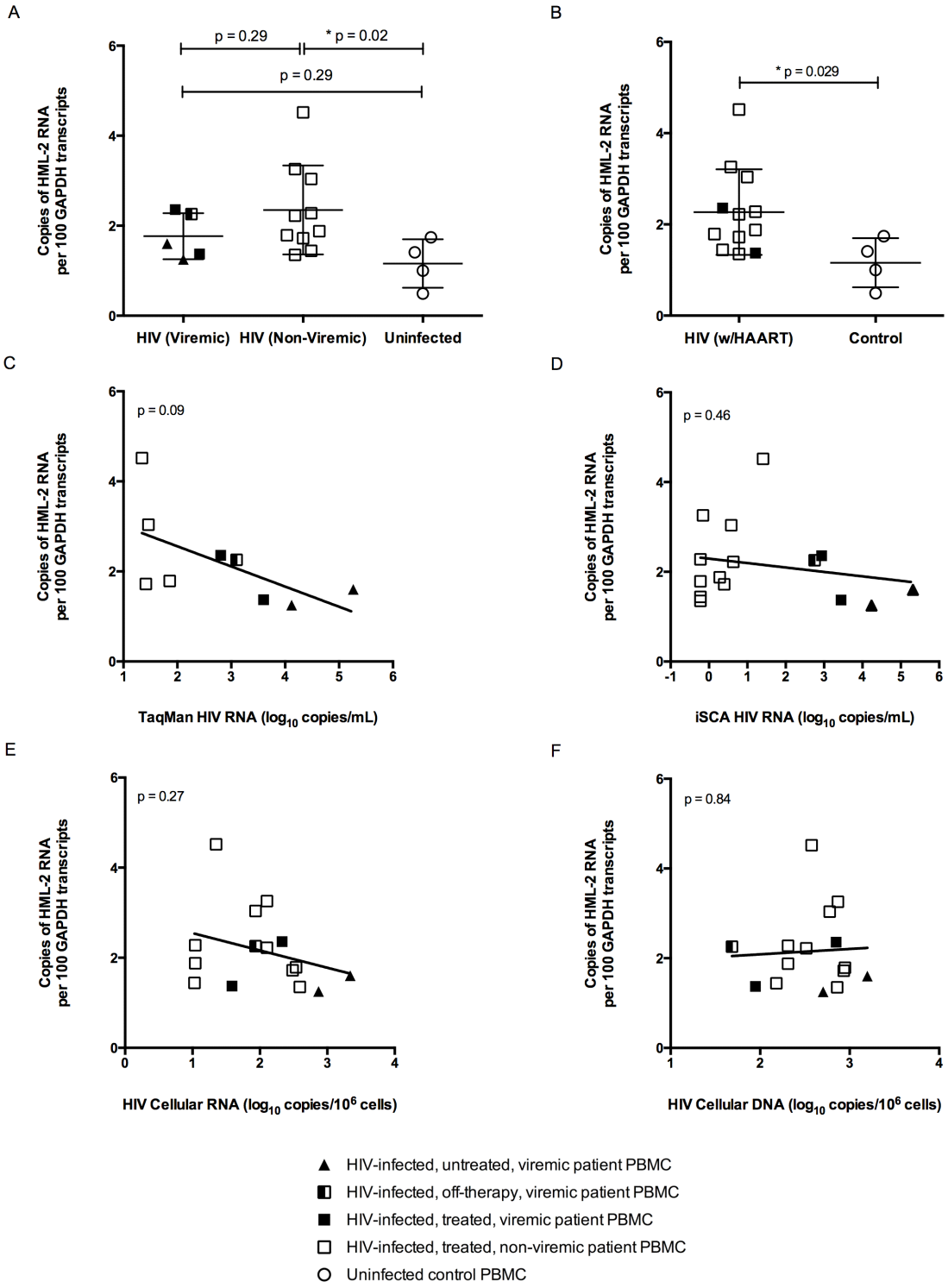
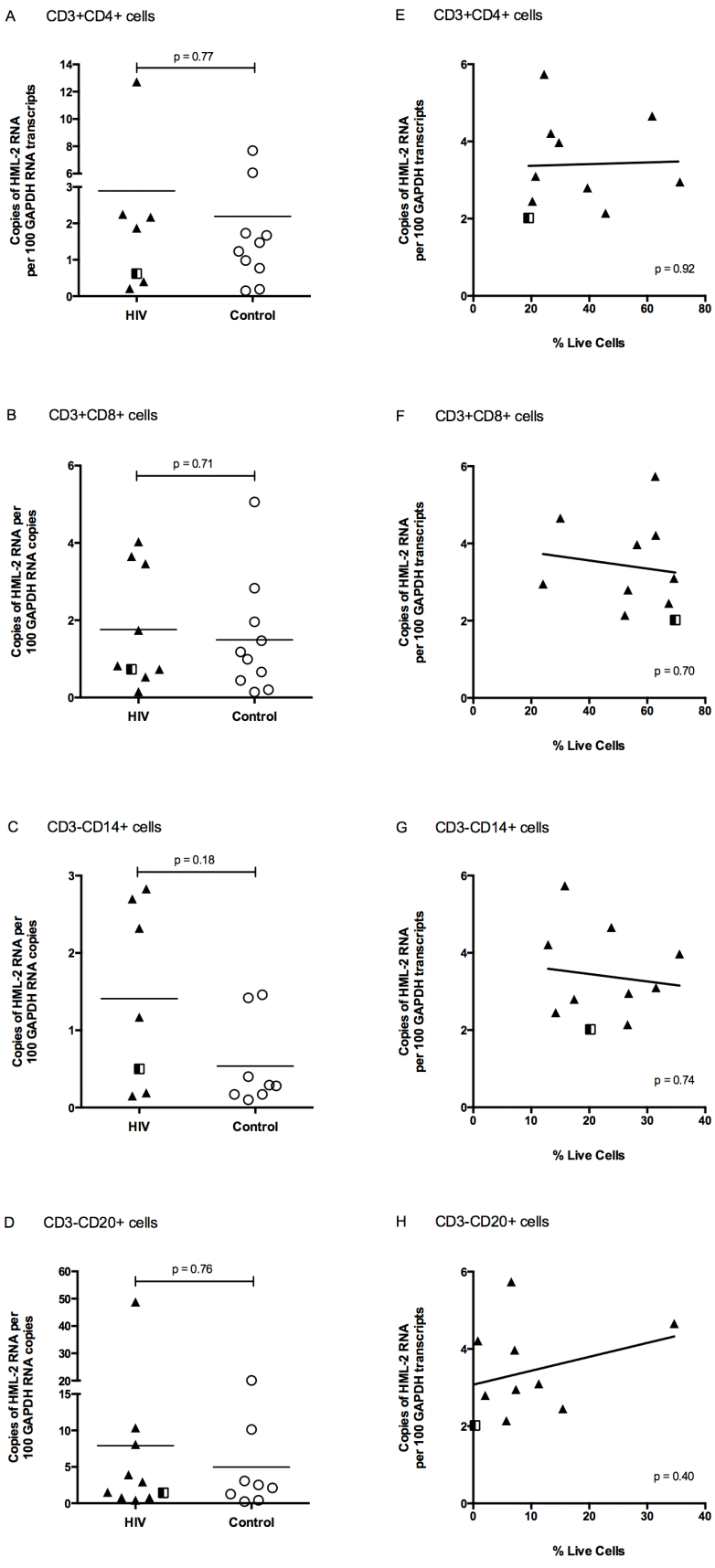

A
p = 0.29
* p = 0.02
p = 0.29
Copies of HML-2 RNA per 100 GAPDH transcripts
HIV (Viremic)
HIV (Non-Viremic)
Uninfected
B
* p = 0.029
Copies of HML-2 RNA per 100 GAPDH transcripts
HIV (w/HAART)
Control
C
p = 0.09
Copies of HML-2 RNA per 100 GAPDH transcripts
TaqMan HIV RNA (log10 copies/mL)
D
p = 0.46
Copies of HML-2 RNA per 100 GAPDH transcripts
iSCA HIV RNA (log10 copies/mL)
E
p = 0.27
Copies of HML-2 RNA per 100 GAPDH transcripts
HIV Cellular RNA (log10 copies/10^6 cells)
F
p = 0.84
Copies of HML-2 RNA per 100 GAPDH transcripts
HIV Cellular DNA (log10 copies/10^6 cells)
HIV-infected, untreated, viremic patient PBMC
HIV-infected, off-therapy, viremic patient PBMC
HIV-infected, treated, viremic patient PBMC
HIV-infected, treated, non-viremic patient PBMC
Uninfected control PBMC

**Figure 3-6. HML-2 RNA expression in different cell types.**

(A-D) PBMCs from viremic HIV-infected patients and healthy controls were stained with antibodies specific to CD3, CD4, CD8, CD14 and CD20. Live cells that were CD3+CD4+, CD3+CD8+, CD3-CD14+ and CD3-CD20+ were sorted for each patient and analyzed for HML-2 RNA expression. (A) CD3+CD4+ T cells, (B) CD3+CD8+ T cells, (C) CD3-CD14+ Monocytes, and (D) CD3-CD20+ B cells. The p values for Mann-Whitney t-tests are shown. (E-H) Correlation between the percent of cell subset (after size and live/dead gating) to HML-2 upregulation in total unsorted PBMCs, including (E) CD3+CD4+ T cells, (F) CD3+CD8+ T cells, (G) CD3-CD14+ Monocytes, and (H) CD3-CD20+ B cells. The p values for linear regression are shown in the lower right hand corner. Refer to the figure legend shown for Fig 3-4.


A
CD3+CD4+ cells
p = 0.77
Copies of HML-2 RNA per 100 GAPDH RNA transcripts
HIV
Control
E
CD3+CD4+ cells
Copies of HML-2 RNA per 100 GAPDH transcripts
p = 0.92
% Live Cells
B
CD3+CD8+ cells
p = 0.71
Copies of HML-2 RNA per 100 GAPDH RNA copies
HIV
Control
F
CD3+CD8+ cells
Copies of HML-2 RNA per 100 GAPDH transcripts
p = 0.70
% Live Cells
C
CD3-CD14+ cells
p = 0.18
Copies of HML-2 RNA per 100 GAPDH RNA copies
HIV
Control
G
CD3-CD14+ cells
Copies of HML-2 RNA per 100 GAPDH transcripts
p = 0.74
% Live Cells
D
CD3-CD20+ cells
p = 0.76
Copies of HML-2 RNA per 100 GAPDH RNA copies
HIV
Control
H
CD3-CD20+ cells
Copies of HML-2 RNA per 100 GAPDH transcripts
p = 0.40
% Live Cells

Taken together, our results imply that HIV-1 infection does not result in the release of a high amount of HML-2 virions in the plasma of patients represented in our cohort, as measured by our assay, contrary to publications showing HML-2 RNA detection at $10^3$ copies/mL and higher [42, 44, 46, 47]. Furthermore, it appears that increased HML-2 expression in HIV-infected patient PBMCs may not be directly reliant upon HIV-1 replication in an individual, exemplified by the lack of association with HIV RNA levels and effect of antiretrovirals on HML-2 expression, and is instead due to an indirect effect of HIV-1 infection. The identities of expressed HML-2 loci in HIV-1 infection remain to be clarified, which will provide insight into pathogenic potential of HML-2 expression in an individual.

## Chapter 4: Use of next-generation sequencing to assess HML-2 proviral expression

Some of the results in this chapter were published previously in:

Bhardwaj N, Montesion M, Roy F, Coffin JM. "Differential expression of HERV-K (HML-2) proviruses in cells and virions of the teratocarcinoma cell line Tera-1." Viruses. 2015;7(3):939-68.

### *4.1 Application of RNASeq to detect HML-2 proviruses*

A potential hurdle to examining the effect of HML-2 expression on the human host is determining which of the multiple HML-2 proviruses are active in different disease states. PCR approaches can reliably detect HML-2 RNA transcripts, however may not be able to discriminate among all the individually expressed HML-2 proviruses. In terms of pathogenic potential and association with disease, the proviral source of HML-2 expression is likely important because of their varying sequence preservation and coding potential [233]. In addition, due to their recent integration, accurate detection of many of the evolutionarily young HML-2 integrations is challenging as they are remarkably similar in sequence and finding unique regions to amplify may not be straightforward for each provirus. Due to sequence similarity, PCR recombination may pose a threat to accurate detection of individual proviruses if more than one is expressed at a time. Gold standard PCR methods like single genome sequencing [184] can effectively circumvent most issues, however amplified targets will be limited by the primer design of the assay and the throughput of the method.

RNASeq provides a high-throughput approach to determine the expression of individual HML-2 loci in the context of cellular genes. Overall, RNASeq is an objective method to approximate total cellular transcription. Total RNA, or just mRNA depending upon the protocol, is converted into cDNA, which is then ligated with 5' and 3' adaptors to allow for amplification of the cDNA sequences prior to deep sequencing and binding of the cDNA to a sequencing flow cell. The resultant reads generated during deep sequencing are either aligned to an appropriate reference genome or transcriptome, or if neither is available, *de novo* assembled into contigs for identification. Importantly, RNASeq does not require that sequences be previously annotated, as reads that do not align to a known reference are still present in the population, which may be useful for the identification of new, unannotated HML-2 integrations, if actively transcribed. In addition, RNASeq can bypass PCR primer bias, important since older HML-2 proviruses that show sequence divergence or lack portions of genomic sequence would still be captured using RNASeq if expressed in a cell, but potentially missed using standard PCR amplification.

RNASeq has been used to quantify expression of specific proviruses belonging to older groups of HERVs, including HERV-H [206] and HERV-W [215], and more recently has been applied to the HML-2 group [26, 87]. A predicted complication in applying RNASeq to HML-2 transcription profiling is caused by the high sequence similarity between recently integrated HML-2 proviruses. Reads originating from highly similar or conserved areas could potentially align to multiple proviruses – these reads are called "multi-reads." The true placement of a multi-read is in question since it may be misaligned to a similar sequence and provide false signal for a related provirus or solo

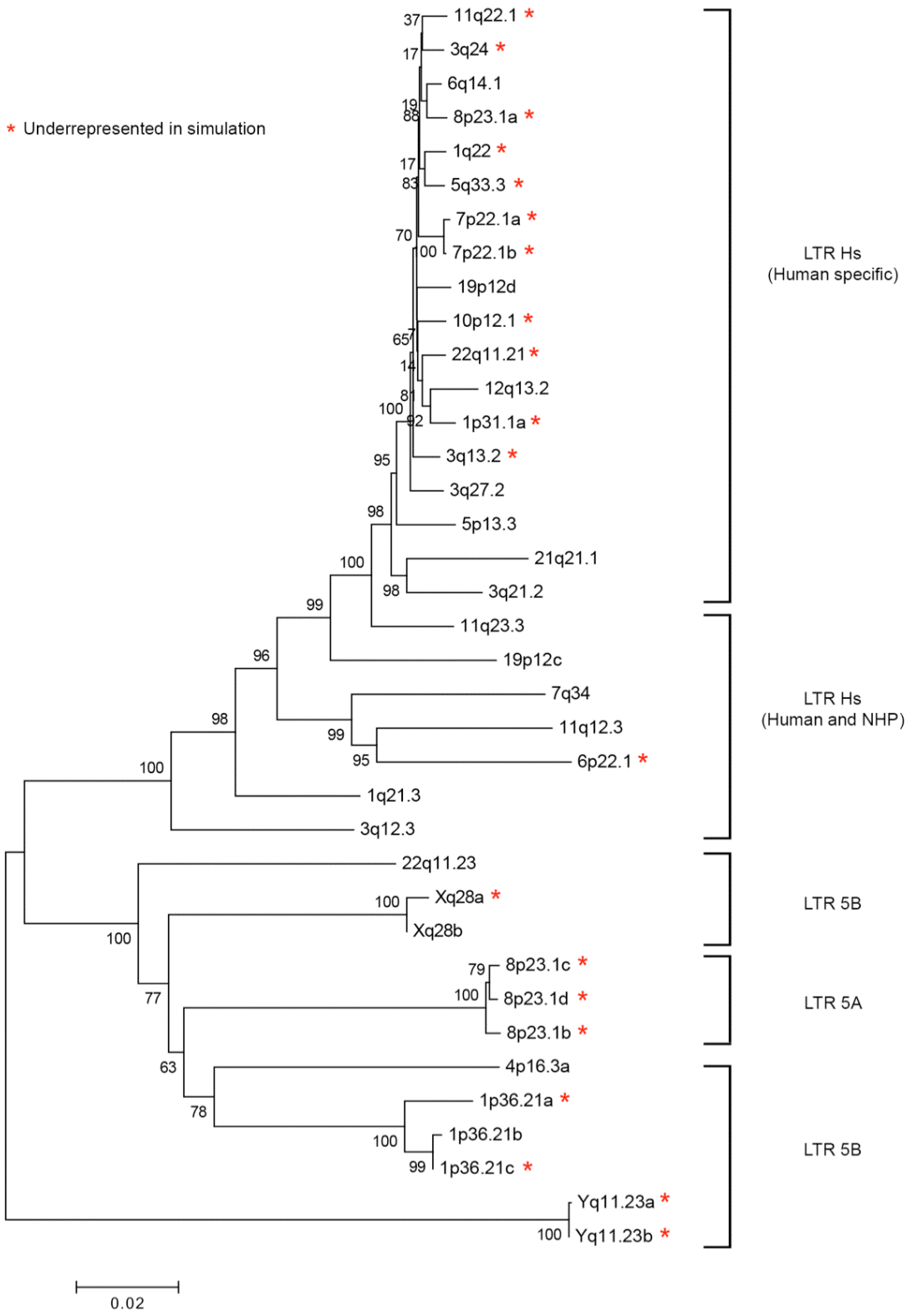
LTR, leading to an inexact representation of provirus transcription. Therefore, to circumvent this problem, only uniquely mapped reads were used for expression analysis, with the reasoning that a truly expressed provirus will also produce reads with unique sequence in addition to more conserved regions that will map to multiple proviruses.

The exclusion of multi-reads improves accuracy, but creates a reporting bias, in which highly similar proviruses could be underrepresented due to a limited number of uniquely aligned reads. To determine the effect of this approach on HML-2 expression analysis, an *in silico* simulation of RNASeq expression analysis was performed. First, RNASeq paired-end reads were simulated based off of the sequences of 93 HML-2 proviruses, generating equal coverage of each proviral sequence. The resultant reads were then aligned back to the HML-2 provirus sequences and filtered to remove the multi-reads that matched more than one provirus. Only the remaining uniquely mapped reads were used as input to calculate gene length normalized RNASeq expression values called FPKM (fragments per kilobase per million mapped reads). The filtering step to remove multi-reads could cause highly similar proviruses to have lower coverage than those with unique sequence, which would be seen as lower FPKM values.

Based on this simulation, we found that, as expected, recently integrated proviruses and duplications appeared to be underrepresented after filtering for unique reads. The proviruses that were negatively affected by >15% (range: 17-86%) are shown on the neighbor-joining phylogenetic tree in Fig 4-1, labeled to show the three main categories of HML-2 proviruses, 5A, 5B, and Hs. All proviruses that appeared to be affected are shown with a red asterisk in future figures to denote their potential underrepresentation.

**Figure 4-1. Phylogenetic tree of underrepresented proviruses in RNASeq.**

A neighbor-joining tree of the underrepresented proviruses (*) identified in our simulation was created using the full provirus sequence, with a total of 10240 positions analyzed. The p-distance method was used to calculate distance and is reported in terms of base differences per site. Bootstrap values are indicated as a percent of 1000 replicates.


* Underrepresented in simulation
11q22.1 *
3q24 *
6q14.1
8p23.1a *
1q22 *
5q33.3 *
7p22.1a *
7p22.1b *
19p12d
10p12.1 *
22q11.21 *
12q13.2
1p31.1a *
3q13.2 *
3q27.2
5p13.3
21q21.1
3q21.2
11q23.3
19p12c
7q34
11q12.3
6p22.1 *
1q21.3
3q12.3
22q11.23
Xq28a *
Xq28b
8p23.1c *
8p23.1d *
8p23.1b *
4p16.3a
1p36.21a *
1p36.21b
1p36.21c *
Yq11.23a *
Yq11.23b *
LTR Hs
(Human specific)
LTR Hs
(Human and NHP)
LTR 5B
LTR 5A
LTR 5B
0.02

Many of the underrepresented loci include the most recent LTR Hs integrations, which have accumulated fewer mutations since their last common ancestors than those resident in the genome for longer periods of time. In addition, proviruses that are known to have arisen by duplication post-integration, including the LTR Hs proviruses on 7p22.1, the LTR 5B proviruses on 1p36.21, Xq28 and Yq11.23, and the LTR 5A proviruses on 8p23.1, are also represented in the RNASeq simulation less frequently than expected, notably the Xq28a/b locus with a ~86% reduction. The LTR Hs provirus 6p22.1, which is not human specific, was also underrepresented in the simulation. Of interest, the two LTR Hs proviruses 3q21.2 and 21q21.1 were shown to have been hypermutated by APOBEC3G [135], accounting for the tight clustering and longer branch length exhibited on the tree. Importantly, all proviruses were detected in the simulation. Therefore, although the true abundance of affected proviruses may be underrepresented, their expression will, nevertheless, be captured in the analysis.

*4.2 HML-2 provirus expression in the teratocarcinoma cell line Tera-1*

With the end goal of applying this RNASeq approach to profile HIV-infected patient samples, this methodology was first validated on the teratocarcinoma cell line Tera-1. Tera-1 cells express HML-2 RNA and protein and are capable of producing HML-2 virions, a phenomenon that has only been reliably identified in a few other cell types [172, 227], though none have been found to be infectious [142]. The biology of this cell line is largely unknown [139, 202], but it has been shown to primarily express HML-2 RNA originating from the provirus at chromosome 22q11.21 and other evolutionarily young integrations [202], and its virions appear to be immature and lacking Env

glycoprotein [16]. By using an RNASeq approach that calculates expression levels based on uniquely aligned reads, we identified a number of distinct HML-2 proviral transcripts expressed in Tera-1 cells, including both evolutionarily older and younger elements.

Two RNASeq libraries were prepared for analysis, with one constructed from Tera-1 cellular RNA and the other from Tera-1 virion RNA. RNASeq reads were clipped to remove poor quality bases, adaptor sequences and reads shorter than 100 bases prior to alignment using the program TopHat2 [121].

Expression of some polymorphic proviruses may not be captured by alignment to a human reference genome because the proviruses may not present in the individual(s) contributing genomic sequence or they were missed in genome assembly. Due to this anticipated issue, reads were aligned to the hg19 build of the human genome as well as to an HML-2 reference genome containing the sequences of 943 solo LTRs, 93 proviruses and a prototype SINE-R element, a type of retrotransposon comprising HERV-K LTR and *env* sequence [177]. Of the 93 proviruses included in the HML-2 reference genome, 4 are present as solo LTRs in hg19 and 2 are present as pre-integration sites. In the HML-2 reference genome, each element was listed as an independent sequence, thus functioning as a catalogue of 1037 HML-2 "chromosomes" during alignment.

Around half of all reads (~47%) that aligned to HML-2 proviruses were multi-reads. Data were either kept in full (referred to as "Unfiltered") or filtered for uniquely aligned reads (referred to as "Unique Only"). Both Unfiltered and Unique Only reads were used to calculate FPKM in order to determine the effect of filtering on expression analysis. To determine FPKM using the Unfiltered reads, the multi-reads present could be assigned proportionally to multiple mapping locations based on the abundance estimations for

each mapping location (e.g. the reads are assigned proportionally to the proviruses with highest number of uniquely aligned reads over those with less, referred to as "Multi-read correct") or in a default manner where multi-reads are assigned to multiple mapping locations uniformly (e.g. if a read maps to 5 locations, each location is assigned 20% of a read). Thus, for FPKM calculation in the Unfiltered read population, both default and Multi-read correct approaches were taken in order to determine relative expression. Another parameter used in FPKM calculation is one that considers whether transcription is occurring in the sense orientation for a particular locus. For the cell library, which was stranded, this parameter was used to estimate abundance using transcripts with the potential for translation into retroviral protein (referred to as "Plus stranded" if performed, and "Unstranded" if not).

The highest expressed HML-2 loci detected in the hg19 alignment were corroborated by the HML-2 alignment. For simplicity, all data presented will reflect the values derived from the hg19 alignment. A comparison of analytical methodologies of the RNASeq data is shown in a heatmap representation in Fig 4-2 A, with red shading marking the highest expressed loci. All discussed proviruses are listed in Table 4-1 with their known aliases and genomic position. In the Unstranded, Unfiltered analysis, many proviruses are noted as expressed. However, in the Unique Only and Multi-read correct analyses, which consider only uniquely aligned reads or reads probabilistically assigned to loci, a dramatic drop off occurs in the FPKM of several of these "expressed" proviruses. Based on this analysis, we only considered proviruses remaining after either Unique Only filtering or Multi-read correction as being reliably expressed in Tera-1 cells and not misaligned to closely related loci. For example, the provirus 8p23.1a (K115; chr8:

7355397-7364859) is not present in Tera-1 cells [202]. Prior to filtering, this provirus was incorrectly assigned 2.5% of all HML-2 reads; after filtering, its expression level dropped to the background value of 0.07%. Of note, the Unique Only and Multi-read correct analyses gave virtually identical results in terms of assigning FPKM to specific proviruses for the Tera-1 data set (Fig 4-2 A). We also analyzed the effect of applying the strandedness option to FPKM assignment, thus considering only reads aligned in the sense orientation of the provirus. Unexpectedly, we found that a number of proviral loci were negatively affected by this distinction. These proviruses are displayed on the heatmap as becoming blue in the final column and include 7q34, 11q12.3 and 11q23.3.

Without considering the strandedness of the read, these proviruses would not have been identified as products of antisense transcription, most likely due to neighboring transcription units. Specifically, in the case of the 7q34 provirus, reads that align to this locus appear to be the product of read-through transcription from the neighboring highly transcribed gene *SSBP1*. For the 11q12.3 provirus, which resides in an intron of the gene *ASRGL1*, aligned reads appear to be result of pre-mRNA present in the total RNA used for library preparation. Finally, at the proviral locus on 11q23.3, aligned reads appear to originate in a HERV-H element located just downstream of the proviral 3' LTR, though there did appear to be an increase in aligned reads throughout the provirus.

To depict how the Unique Only analysis affected provirus representation in Tera-1 cells, Fig 4-2 B shows how the estimated age of integration for expressed proviruses changed between an Unfiltered, Plus stranded alignment and the Unique Only, Plus stranded alignment. Recently integrated proviruses are still represented in analysis,

however they are 2/3 less abundant, leading to a perceived overrepresentation of older elements that carry more unique sequence.

**Figure 4-2. RNASeq analysis of HML-2 expression in Tera-1 cells.**

(A) RNASeq reads derived from Tera-1 cellular RNA were aligned to the hg19 build of the human genome, using either a stranded ("Plus Stranded") or unstranded ("Unstranded") alignment. Aligned reads were either kept in full ("Unfiltered"), or were filtered based on mapping quality scores to only retain reads that uniquely aligned to one map location ("Unique Only"). The fragments per kilobase per million mapped reads (FPKM) values representing relative expression in Tera-1 cells were determined either with a multi-read correction parameter ("Multi-read Correct") that proportionally allocates multi-reads to mapping locations, or without this parameter. FPKM values for selected HML-2 proviruses and the cellular genes *GAPDH* and *ACTB* across the analyses were log-normalized and used for heatmap generation to demonstrate the effects of the different analyses on expression levels. Proviruses and gene loci are divided into four groups according to their relative values following the different analyses: stable (Group 1); decrease after Unique Only (Group 2); decrease after Plus stranded alignment (Group 3); and decrease after Unique Only and Plus stranded analysis (Group 4). Log-normalized FPKM is shown by the colors from high (red) to low (blue), as indicated in the chart to the right. The (*) symbols refer to proviruses predicted to be underrepresented by 15% or more based on the *in silico* simulation (B) The abundance of transcripts after the Plus stranded, Unfiltered and the Plus Stranded, Unique Only analyses are plotted against estimated times of integration to show the effect of the Unique Only analysis on recently integrated proviruses. The 0-2 mya group includes human specific integrations with high

sequence similarity predicted to be underrepresented in the Unique Only RNASeq *in silico* simulation. The relative abundance in Tera-1 cells was calculated for each provirus based on (provirus FPKM)/(total HML-2 provirus FPKM) x 100. Elements without 5' or 3' LTRs were unsuitable for age estimation and are not included.

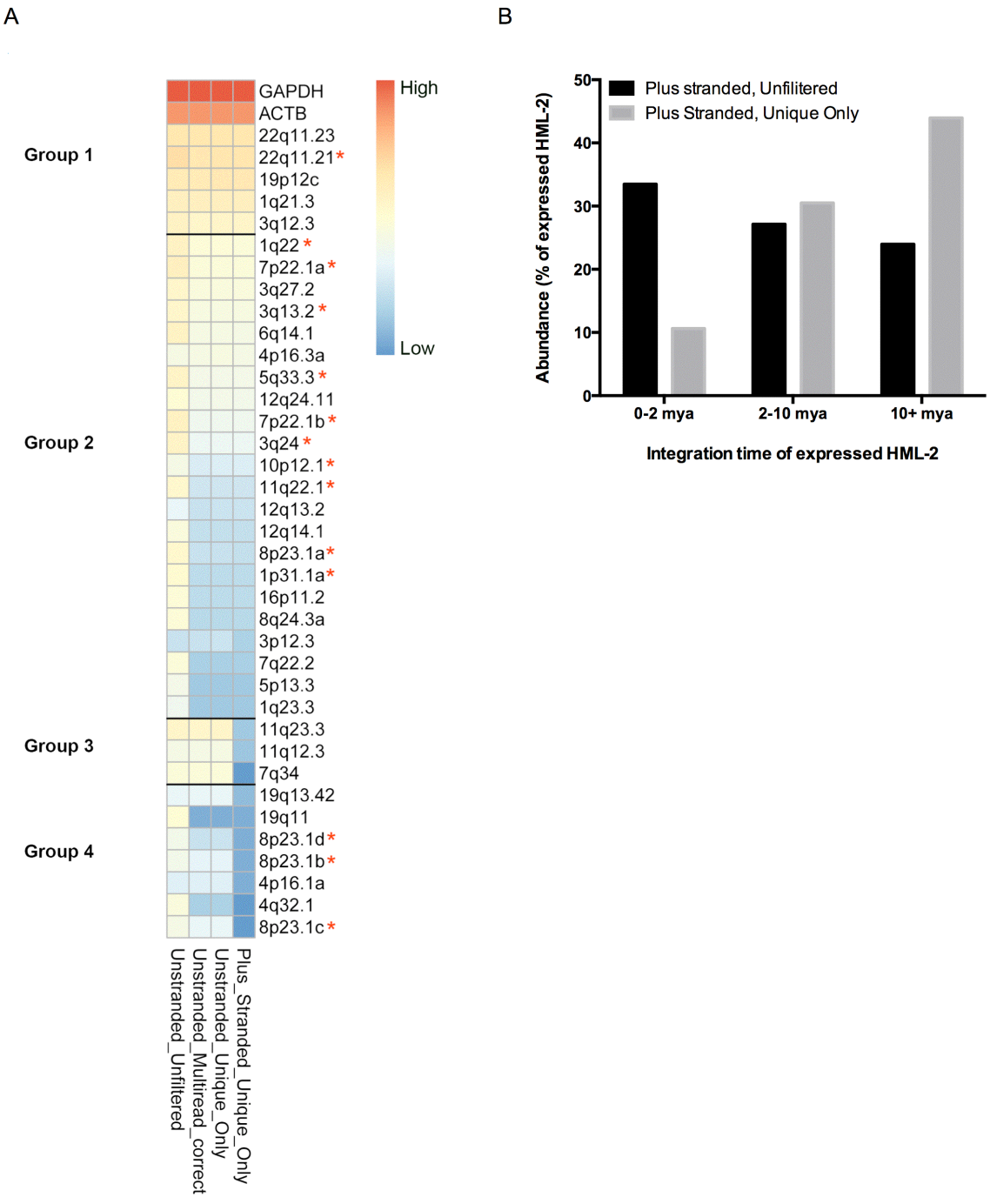


**Table 4-1. Names and locations for HML-2 proviruses discussed in Chapter 4.**

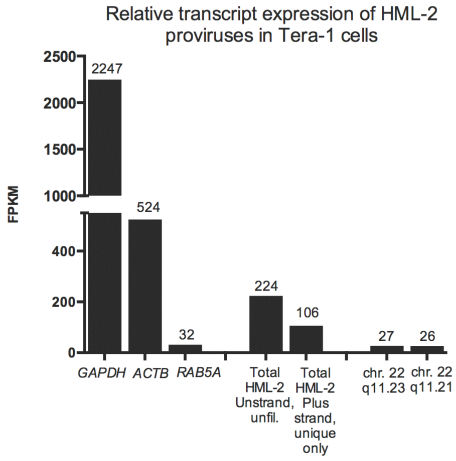
| Provirus | Alias | Chromosomal Location (hg19) |
|---|---|---|
| **1p31.1a** | K4, K116, ERV-K1 | chr1: 75842771-75849143 |
| **1p36.21a** | N/A | chr1: 12840260-12846364 |
| **1p36.21b** | K(OLDAL023753), K6, K76 | chr1: 13458305-13467826 |
| **1p36.21c** | K6, K76 | chr1: 13678850-13688242 |
| **1q21.3** | N/A | chr1: 150605284-150608361 |
| **1q22** | K102, K(C1b), K50a, ERVK-7 | chr1: 155596457-155605636 |
| **1q23.3** | K110, K18, 1 (+) K(C1a), ERVK-18 | chr1: 160660575-160669806 |
| **3q12.3** | K(II), ERVK-5 | chr3: 101410737-101419859 |
| **3q13.2** | K106, K(C3), K68, ERVK-3 | chr3: 112743123-112752282 |
| **3q21.2** | K(I), ERVK-4 | chr3: 125609302-125618416 |
| **3q24** | ERVK-13 | chr3: 148281477-148285396 |
| **3q27.2** | K50b, K117, 3 (-) ERVK-11 | chr3: 185280336-185289515 |
| **4p16.3a** | N/A | chr4: 234989-239459 |
| **5q33.3** | K107/K10, K(C5), ERVK-10 | chr5: 156084717-156093896 |
| **6p22.1** | K(OLDAL121932), K69, K20 | chr6: 28650367-28660735 |
| **6q14.1** | K109, K(C6), ERVK-9 | chr6: 78427019-78436083 |
| **6q25.1** | N/A | chr6: 151180749-151183574 |
| **7p22.1a** | K108L, K(HML.2-HOM), K(C7), ERVK-6 | chr7: 4622057-4631528 |
| **7p22.1b** | K108R, ERVK-6 | chr7: 4630561-4640031 |
| **7q34** | K(OLDAC004979), ERVK-15 | chr7: 141450926-141455903 |
| **8p23.1a** | K115, ERVK-8 | chr8: 7355397-7364859 |
| **8p23.1b** | K27 | chr8: 8054700-8064221 |
| **8p23.1c** | N/A | chr8: 12073970-12083497 |
| **8p23.1d** | KOLD130352 | chr8: 12316492-12326007 |
| **10p12.1** | K103, K(C10) | chr10: 27182399-27183380 |
| **10p14** | K(C11a), K33, ERVK-16 | chr10: 6867109-6874635 |
| **11q12.3** | K(OLDAC004127) | chr11: 62135963-62150563 |
| **11q22.1** | K(C11c), K36, K118, ERVK-25 | chr11: 101565794-101575259 |
| **11q23.3** | K(C11b), K37, ERVK-20 | chr11: 118591724-118600883 |
| **12q13.2** | N/A | chr12: 55727215-55728183 |
| **12q14.1** | K(C12), K41, K119, ERVK-21 | chr12: 58721242-58730698 |
| **12q24.11** | N/A | chr12: 111007843-111009325 |
| **12q24.33** | K42 | chr12: 133667120-133673132 |

| | | |
|---|---|---|
| **14q11.2** | K(OLDAL136419), K71 | chr14: 24480625-24484121 |
| **16p13.3** | K(OLDAC004034) | chr16: 2976160-2077661 |
| **19p12a** | K52 | chr19: 20387400-20397512 |
| **19p12b** | K113 | chr19: 21841536-21841542 (empty site) |
| **19p12c** | K51 | chr19: 22757824-22764561 |
| **19p12d** | N/A | chr19: 22414379-22414382 (empty site) |
| **19q13.12a** | N/A | chr19: 36063207-36067434 |
| **19q13.12b** | K(OLDAC012309), KOLD12309 | chr19: 37597549-37607066 |
| **19q13.41** | N/A | chr19: 53248274-53252591 |
| **19q13.42** | LTR13 | chr19: 53862348-53868044 |
| **21q21.1** | K60, ERVK-23 | chr21: 19933916-19941962 |
| **22q11.21** | K101, K(C22), ERVK-24 | chr22: 18926187-18935307 |
| **22q11.23** | K(OLDAP000345), KOLD345 | chr22: 23879930-23890615 |
| **Xq28a** | K63 | chrX: 153817163-153819562 |
| **Xq28b** | K63 | chrX: 153836675-153844015 |
| **Yq11.23a** | N/A | chrY: 26397837-26401035 |
| **Yq11.23b** | N/A | chrY: 27561402-27564601 |

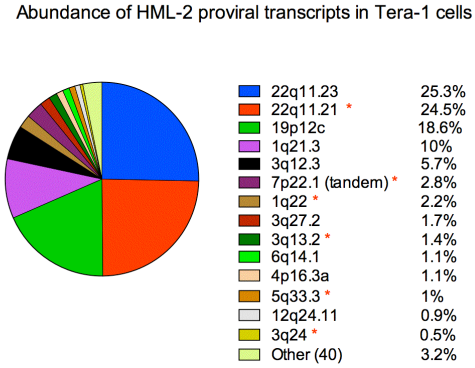
**Figure 4-3. HML-2 expression in Tera-1 cells and virions.**

(A-B) RNASeq reads originating from Tera-1 cells were aligned to the hg19 build of the human genome and analyzed using the Plus stranded, Unique Only analysis, except as indicated. (E-F) RNASeq reads originating from Tera-1 virions were aligned to the hg19 build of the human genome and analyzed using the Unstranded, Unique Only analysis, except as indicated, due to the input library not being stranded. (A, E) Relative transcript expression values (FPKM) for cellular genes, total HML-2 and the most abundantly expressed or packaged HML-2 transcripts are plotted for Tera-1 cells (A) and Tera-1 virions (E). (B, F) Abundance of transcripts for each provirus in Tera-1 cells (B) and virions (F) is plotted according to (provirus FPKM)/(total HML-2 FPKM) x 100. Proviruses with (*) were predicted to be underrepresented by the *in silico* analysis, as used in Figure 4-1. (C) Open reading frames for *gag*, *pro*, *pol* and *env* were determined for proviruses making up 96.8% of all HML-2 reads shown in Fig 4-3 B. The almost full-length *gag* ORF from 22q11.23 was included in this analysis. If a provirus had the potential to express full ORF(s), the abundance of the provirus in the cell was allocated to each ORF, to represent the maximum probability of that ORF being expressed. Splicing was not considered for this analysis. (D) Type 1/2 status was determined for HML-2 proviruses making up 96.8% of all HML-2 reads, listed in Fig 4-3 B. Unknown indicates that the *pol-env* boundary region was not present in the provirus, preventing identification of provirus type.

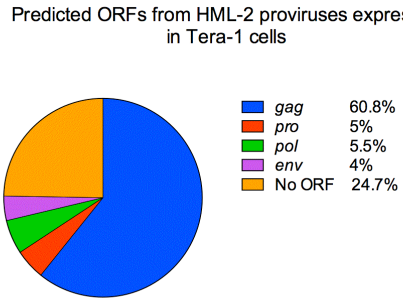
A


Relative transcript expression of HML-2 proviruses in Tera-1 cells
FPKM
2500
2000
1500
1000
400
200
0
2247
524
32
224
106
27
26
GAPDH
ACTB
RAB5A
Total HML-2 Unstrand, unfil.
Total HML-2 Plus strand, unique only
chr. 22 q11.23
chr. 22 q11.21

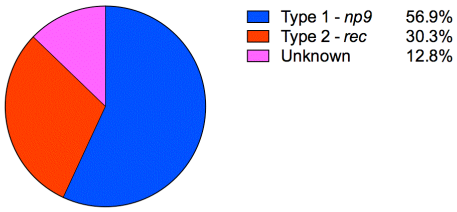
B


Abundance of HML-2 proviral transcripts in Tera-1 cells
22q11.23 25.3%
22q11.21 * 24.5%
19p12c 18.6%
1q21.3 10%
3q12.3 5.7%
7p22.1 (tandem) * 2.8%
1q22 * 2.2%
3q27.2 1.7%
3q13.2 * 1.4%
6q14.1 1.1%
4p16.3a 1.1%
5q33.3 * 1%
12q24.11 0.9%
3q24 * 0.5%
Other (40) 3.2%

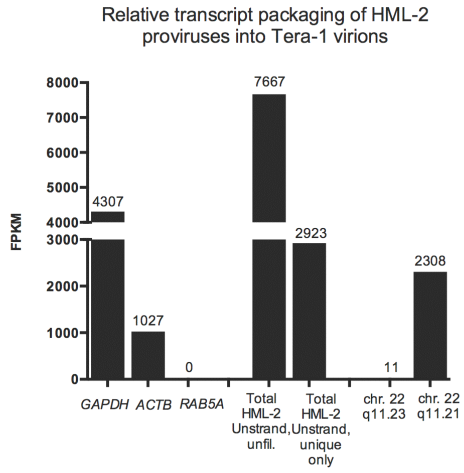
C


Predicted ORFs from HML-2 proviruses expressed in Tera-1 cells
gag 60.8%
pro 5%
pol 5.5%
env 4%
No ORF 24.7%

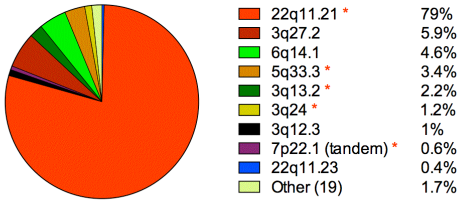
D


HML-2 provirus expression by type in Tera-1 cells
Type 1 - np9 56.9%
Type 2 - rec 30.3%
Unknown 12.8%

E


Relative transcript packaging of HML-2 proviruses into Tera-1 virions
FPKM
8000
7000
6000
5000
4000
3000
2000
1000
0
4307
1027
0
7667
2923
11
2308
GAPDH
ACTB
RAB5A
Total HML-2 Unstrand, unfil.
Total HML-2 Unstrand, unique only
chr. 22 q11.23
chr. 22 q11.21

F


Abundance of HML-2 proviral transcripts in Tera-1 virions
22q11.21 * 79%
3q27.2 5.9%
6q14.1 4.6%
5q33.3 * 3.4%
3q13.2 * 2.2%
3q24 * 1.2%
3q12.3 1%
7p22.1 (tandem) * 0.6%
22q11.23 0.4%
Other (19) 1.7%

*4.3 Relationship between HML-2 provirus expression in cells and packaging into virions in the Tera-1 cell line*

Using the Plus stranded, Unique Only approach, HML-2 transcription in Tera-1 cells was compared to the cellular genes *GAPDH*, *ACTB* and *RAB5A* (Fig 4-3 A). The Unique Only analysis removed roughly half of all HML-2 reads present in the unfiltered alignment, yet total expression of this group was still readily quantifiable, at ~1/200$^{th}$ the level of the metabolic gene *GAPDH* and ~1/5$^{th}$ the level of the cytoskeletal gene *ACTB* (β-actin). The top two expressed HML-2 proviruses, the LTR 5B provirus at 22q11.23 and the LTR Hs provirus at 22q11.21 (K101, Fig 4-3 A-B), were each detected at a level comparable to that of the cellular gene *RAB5A*, which encodes a protein localized on early endosomes, and together made up roughly half of all the HML-2 reads generated from Tera-1 cells.

LTR Hs type proviruses were the most commonly expressed proviruses in Tera-1 cells (12 out of the top 14, Fig 4-3 B), and included 7 human specific integrations that were likely to be underrepresented (indicated with red asterisks as in Fig 4-1 and elsewhere). The tandem duplicated LTR Hs proviruses on chromosome 7p22.1 (K108) [233] were considered together since they are nearly identical in sequence and reads could have originated from either of them. Interestingly, two ancient LTR 5B proviruses, on chromosomes 22q11.23 and 4p16.3a, were also expressed in the cells.

The ORFs for the genes *gag*, *pro, pol* and *env* vary among the HML-2 proviruses. To assess the contribution of the identified proviruses to virion production, we calculated the relative numbers of transcripts belonging to proviruses capable of potentially expressing full-length gene products. We found that most (61%) expressed HML-2 sequence was

capable of encoding *gag*, largely from the *gag* ORFs present on proviruses 22q11.23 and 22q11.21 (Fig 4-3 C). Of interest, the Gag protein encoded by 22q11.21 appears to be full length (666 amino acids) but the Gag encoded by 22q11.23, which is a more ancient provirus, is predicted to be truncated by 43 amino acids at the C-terminus. The *pro* (5%), *pol* (5.5%) and *env* (4%) ORFs were much less well represented, and a significant fraction (24.7%) of the expressed HML-2 sequence was derived from proviruses that lack coding capability altogether (Fig 4-3 C). The provirus at 22q11.21 encodes *pro* but has an early stop codon leading to a major C-terminal truncation (271AA instead of 334AA), thus it was not counted towards the available *pro* ORF. The majority of HML-2 proviruses expressed were Type 1 (Fig 4-3 D), which is typified by a 292-bp deletion at the *pol-env* boundary, resulting in a non-fusogenic Env, and encodes the accessory gene *np9* [2]. Type 2 proviruses, which made up ~30% of expressed HML-2 proviruses, retain full sequence at the *pol-env* boundary and encode the accessory gene *rec* [143].

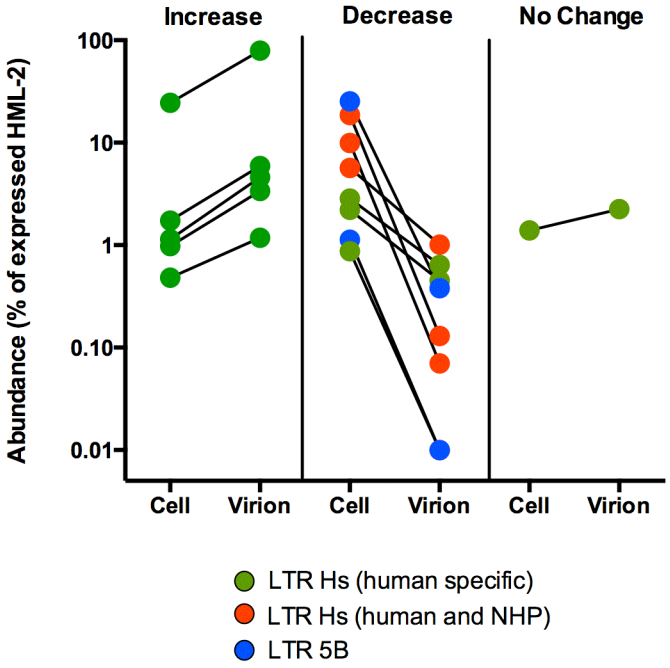
In reads generated from Tera-1 virions, HML-2 sequences were more frequently represented compared to their detection in the cells, as expected, exhibiting >25-fold increase in FPKM (Fig 4-3 E). Virions also appeared to non-specifically package the highly expressed cellular mRNAs from *GAPDH* and *ACTB*, which were increased about 2-fold in FPKM from their levels in cells, but not *RAB5A*, which was not detected in the virions (Fig 4-3 E).

The virion reads aligned primarily to the type 1 provirus on chromosome 22q11.21, making up ~79% of all HML-2 reads (Fig 4-3 E-F). This observation is in agreement with a previous publication assessing the origins of packaged HML-2 RNA from Tera-1 virions [202]. In virions, over 90% of packaged genomes originated from Env-defective

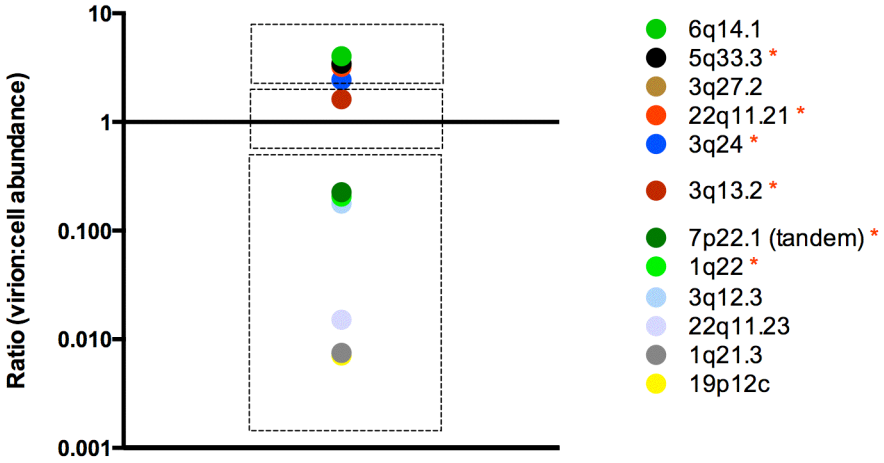
**Figure 4-4. HML-2 packaging shows preference for recently integrated proviruses.**

(A) The abundance of proviruses expressed in the cell and packaged into virions was calculated as described in Figure 4-3. These values were plotted side-by-side to show an increased abundance (panel 1, left), decreased abundance (panel 2, middle) or similar abundance (panel 3, right) for proviruses packaged in virions as compared to their expression in the cell. LTR types of proviruses detected are indicated, with LTR Hs (human specific) in green, LTR Hs (in humans and non-human primates) in red and LTR 5B in blue. Two proviruses (12q24.11 and 4p16.3a) that were not detected in virions were plotted at 0.01% in panel 2. (B) The identities of the proviruses and the ratios of their virion to cell abundance are shown. Proviruses with (*) were predicted to be underrepresented by the *in silico* analysis (Figure 4-1).

A


Observed changes in the abundance of HML-2 transcripts detected in Tera-1 virions as compared to cells
Increase
Decrease
No Change
Abundance (% of expressed HML-2)
100
10
1
0.10
0.01
Cell
Virion
Cell
Virion
Cell
Virion
LTR Hs (human specific)
LTR Hs (human and NHP)
LTR 5B

B


Ratio (virion:cell abundance)
10
1
0.100
0.010
0.001
6q14.1
5q33.3 *
3q27.2
22q11.21 *
3q24 *
3q13.2 *
7p22.1 (tandem) *
1q22 *
3q12.3
22q11.23
1q21.3
19p12c

Type 1 proviruses, mainly due to the abundance of the 22q11.21 transcripts. The top 6 packaged transcripts are all members of the human specific LTR Hs group, a major distinction from the proviruses expressed in Tera-1 cells, which included LTR Hs proviruses that were not human specific as well as older LTR 5B proviruses (Fig 4-3 B).

HML-2 transcripts packaged in Tera-1 virions were more abundant, less abundant or present in roughly equal proportion to their expression in cells (Fig 4-4 A), and this pattern reflected the relative time of their integration. The transcripts with the highest increase in abundance in virions were all derived from recently integrated human specific LTR Hs proviruses (Fig 4-4 B), some of which were predicted to be underrepresented in the analysis (Fig 4-1), noted with red asterisks as before, whereas those that decreased in abundance mostly originated from either older LTR Hs or LTR 5B proviruses. For example, transcripts from the LTR 5B provirus at 22q11.23 made up 25.3% of all cellular HML-2 reads, but its abundance in the virions was 0.4%, a 63-fold decrease (Fig 4-4 B). Transcripts from the LTR 5B provirus at chromosome 4p16.3a, expressed in cells at about 1%, were not even detected in the virions. Other transcripts with large decreases, on chromosomes 1q21.3, 3q12.3 and 19p12, were derived from older LTR Hs integrations.

*4.4 HML-2 proviruses are transcribed through a variety of mechanisms*

To determine the relatedness of the expressed proviruses shown in Fig 4-3 B, the relationship of their 5' LTRs was visualized using a neighbor-joining tree [241] (Fig 4-5 A). As expected, the recently integrated (human-specific) LTR Hs proviruses clustered very closely and for the most part could not be definitively assigned to branches due to their similarity, as seen by the low bootstrap support values generated (Fig 4-5 A).

However, the relationship of the older LTR Hs elements and LTR 5B elements could be ascertained from the tree and was clearly distinct from the recent LTR Hs integrations. The association of divergent LTR types with transcribed proviruses in the Tera-1 cells implies either that the promoter elements of these distinct LTRs were all functional, or that there are alternative ways (i.e. 5' LTR independent) in which some of the older more mutated proviruses were transcribed.

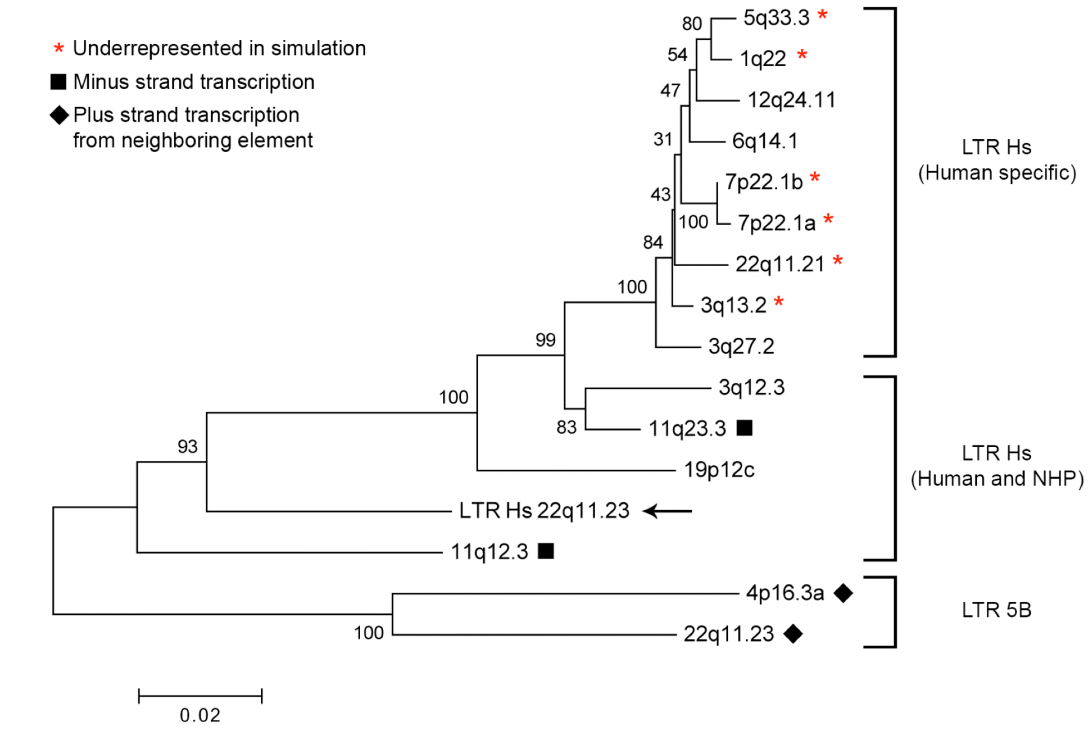
Visualization of HML-2 reads aligned to their map locations using the UCSC Genome Browser [120] or the Integrative Genomics Viewer [249] can inform whether transcription of a provirus is driven by its 5' LTR. That is, 5' LTR driven proviral transcription should start within the provirus, whereas transcription caused by read-through from a neighboring transcription unit results in reads aligning to the provirus as well as flanking sequence intermediate to the transcriptional start and/or end. Transcription driven from a neighboring element may also result in minus strand reads if the provirus and element are in opposite transcriptional orientation, a phenomenon relevant to LTR Hs proviruses 7q34, 11q12.3 and 11q23.3 (Fig 4-2 A and 4-5 A, solid squares).

Transcription of the ancient LTR 5B proviruses 4p16.3a and 22q11.23 (Fig 4-5 A, diamonds) appears to be driven by sequences other than the corresponding 5' LTR. Provirus 4p16.3a (FPKM = 1.19) resides in an intron for the expressed gene *ZNF876P* (FPKM = 9.45). Reads align evenly to *ZNF876P* pre-mRNA intronic sequence, which includes the provirus and surrounding sequence. This result implies that the provirus is not being specifically transcribed; rather, it is preserved in an incompletely removed

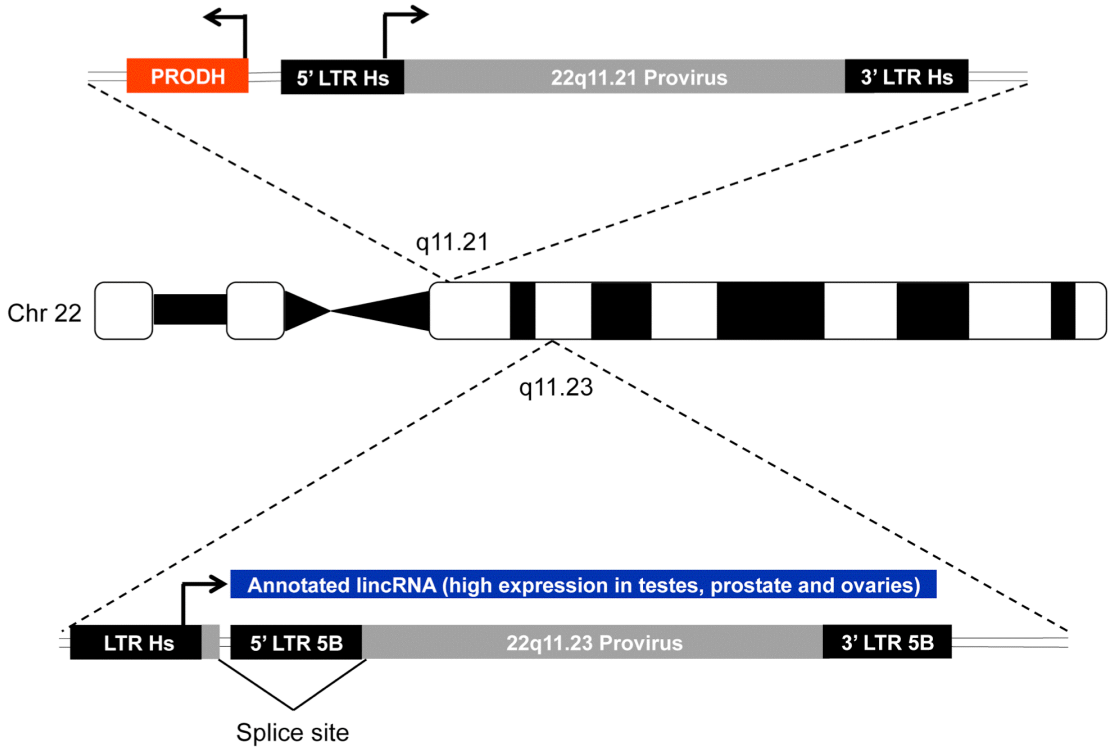
**Figure 4-5. Transcription of HML-2 proviruses is driven by the native LTR or a nearby element.**

(A) Neighbor-joining tree of the 5' LTR sequences of the HML-2 proviruses expressed in Tera-1 cells. The p-distance method was used to calculate distance and bootstrap values are indicated (1000 replicates). Proviruses with (*) were predicted to be underrepresented by the *in silico* analysis, as in Figure 4-1. Solid squares indicate those proviruses (11q23.3 and 11q12.3) with minus strand transcription driven by a neighboring element. Solid diamonds indicate those proviruses (4p16.3a and 22q11.23) with plus strand transcription driven by a neighboring element and not the proviral 5' LTR. (B) A cartoon of the top two expressed proviruses, both located on chromosome 22, and their method of transcription. Provirus 22q11.21 (LTR Hs, FPKM=26.11) is located 2.1kb downstream from the expressed gene *PRODH* (Proline Dehydrogenase (oxidase) 1, FPKM = 11.53) but in the opposite transcriptional orientation. The 5' LTR of 22q11.21 appears to drive proviral transcription in Tera-1 cells. Provirus 22q11.23 (FPKM = 26.94) appears to be transcribed solely through the use of an LTR Hs (FPKM = 0.31) located 551bp upstream from the provirus. This transcript coincides with an annotated lincRNA [30].

A


* Underrepresented in simulation
■ Minus strand transcription
◆ Plus strand transcription from neighboring element
80
54
47
31
43
100
84
100
99
100
83
93
100
5q33.3 *
1q22 *
12q24.11
6q14.1
7p22.1b *
7p22.1a *
22q11.21 *
3q13.2 *
3q27.2
3q12.3
11q23.3 ■
19p12c
LTR Hs 22q11.23
11q12.3 ■
4p16.3a ◆
22q11.23 ◆
LTR Hs (Human specific)
LTR Hs (Human and NHP)
LTR 5B
0.02

B


PRODH
5' LTR Hs
22q11.21 Provirus
3' LTR Hs
q11.21
Chr 22
q11.23
Annotated lincRNA (high expression in testes, prostate and ovaries)
LTR Hs
5' LTR 5B
22q11.23 Provirus
3' LTR 5B
Splice site

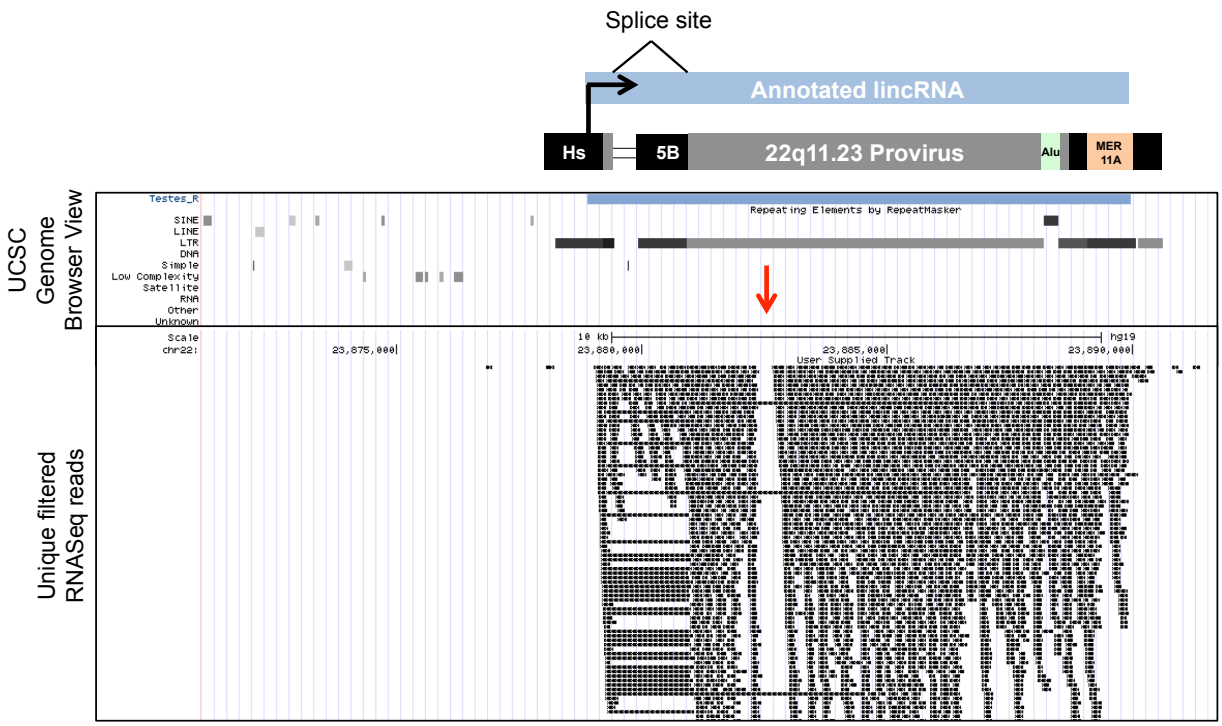
intron. Remarkably, visualization of the highly expressed LTR 5B 22q11.23 provirus (FPKM = 26.94) revealed a fragmentary LTR Hs element (FPKM = 0.31) 551bp upstream, which appeared to be the start site for some fraction of 22q11.23 proviral transcription (Fig 4-5 B, Fig 4-6). Transcription appeared to start midway through the R region of the upstream LTR Hs at position 826. Splicing of the transcript occurred at position 1074 (*gag* leader) of the LTR Hs element into position 1018 (*gag* leader) of the LTR 5B provirus and followed the GU-AG rule (Fig 4-5 B, Fig 4-6). Interestingly, this spliced transcript has been annotated as a lincRNA (TCONS_l2_00017644), though its function is unknown. Though the majority of reads align to the upstream LTR Hs element, indicating transcriptional activity, there are also reads aligned to the 22q11.23 proviral 5' LTR 5B, potentially indicating that its native promoter is active. The relative contributions of each LTR to promoting transcription of the 22q11.23 provirus are not immediately clear based on the alignment alone. Of note, the expression of the upstream LTR Hs appeared to be artificially low since reads primarily aligned only to a small region at the end of the element, affecting the FPKM calculation. The relationship of the 22q11.23 LTR Hs sequence to that of other expressed LTR Hs sequences is shown in Fig 4-5 (black arrow).

The read-through transcription that appears to be driving expression of the LTR 5B proviruses can be contrasted with the clearly 5' LTR driven transcription of the top expressed LTR Hs provirus, 22q11.21 (FPKM = 26.1) (Fig 4-3 A-B, Fig 4-5 B, Fig 4-7). This provirus is integrated 2.1kb downstream from the transcriptional start of the expressed cellular gene *PRODH* (FPKM = 11.53). Their transcriptional orientations are divergent (Fig 4-5 B, Fig 4-7), although their expression has been reported to be linked

[234]. Due to the sequence similarity of this provirus with other recently integrated LTR Hs proviruses, some internal and LTR regions do not show coverage after the Unique Only filter is applied (Fig 4-7). The transcriptional start for this provirus appears to occur around and after position 780 on the 5' LTR, near or at the expected site on the U3-R border at position 793. In support of the role of the 5' LTR in driving proviral transcription, there are only a few reads aligned to the flanking region upstream of the provirus, indicating that upstream elements do not contribute to provirus transcription.

**Figure 4-6. UCSC genome browser view of reads aligned to provirus 22q11.23.**

RNASeq reads from Tera-1 cells were aligned to the hg19 build of the human genome, filtered for Unique Only reads and a BEDfile of the alignment was uploaded to the UCSC genome browser as a custom track. A screenshot of the UCSC genome browser around LTR 5B provirus 22q11.23 (chr22: 23879930-23890615) is shown. A cartoon is placed above the UCSC Genome browser view to clearly indicate: (1) the LTR Hs element driving transcription; (2) the 22q11.23 LTR 5B provirus; (3) the splice site observed between the LTR Hs element and the provirus; and (4) the lincRNA annotated in the genome that overlaps this locus. The black arrow indicates the start site and direction of transcription observed in Tera-1 cells and the red arrow points to a region of the provirus that was not covered after the Unique Only filtering. LTR sequence is shown in black and the internal genomic sequence (*gag-pro-pol-env*) is shown in gray. The UCSC Genome browser view shows the Repeat Masker annotations and also the lincRNA highly expressed in the testes (lincRNA accession number: TCONS_l2_00017644). Of note, an AluY element is inserted at the end of *env* in the 22q11.23 provirus and a MER11A element inserted into its 3' LTR appears to be the site of transcriptional termination. A key depicting the symbols used for reads crossing splice junctions and reads that do not is shown below the screenshot.



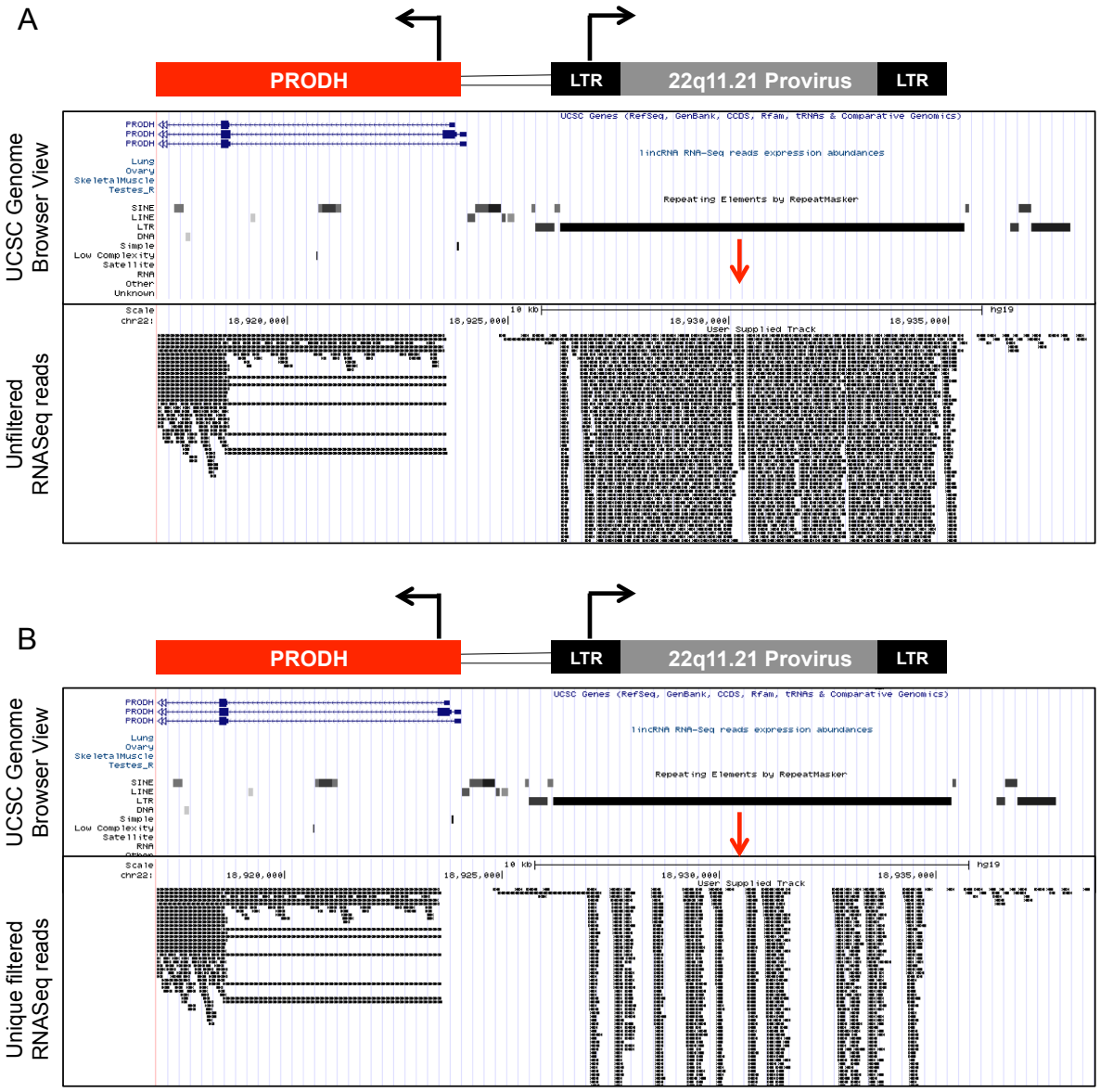

KEY:

Reads crossing splice junctions
(Note: appear as one line of uniform boxes spanning 100s-1000s nt)

Reads that do not cross splice junctions
(Note: appear as a grouping of 1-3 boxes and span 100-301 nt)

**Figure 4-7. UCSC genome browser view of reads aligned to provirus 22q11.21 before and after Unique Only filtering.**

RNASeq reads from Tera-1 cells were aligned to the hg19 build of the human genome, and were left Unfiltered (A) or filtered for Unique Only reads (B). BEDfiles of the alignments were uploaded to the UCSC genome browser as custom tracks. Screenshots of the UCSC genome browser around LTR Hs provirus 22q11.21 (chr22: 18926187-18935307) are shown. A cartoon is placed above the UCSC Genome browser view to indicate that expression of the 22q11.21 provirus is distinct from the nearby PRODH gene. The black arrow indicates the start sites and direction of transcription for 22q11.21 and *PRODH* observed in Tera-1 cells. The red arrow points to a region of the provirus that exhibited poor read coverage in the Unfiltered and Unique Only alignments. LTR sequence is shown in black and the internal genomic sequence (*gag-pro-pol-env*) is shown in gray. The UCSC Genome browser view shows the Repeat Masker annotations and the transcript annotation for the gene *PRODH*. A key depicting the symbols used for reads crossing splice junctions and reads that do not is shown below the screenshots.




KEY:

Reads crossing splice junctions
(Note: appear as one line of uniform boxes spanning 100s-1000s nt)

Reads that do not cross splice junctions
(Note: appear as a grouping of 1-3 boxes and span 100-301 nt)

*The experiments described below were performed by Meagan Montesion and supervised by Neeru Bhardwaj.*

Retroviral 5' LTRs generally possess all promoter elements necessary to drive the transcription of associated viral genes [109, 154, 200]. However, other factors in cells, such as epigenetic effects and expression of nearby genes, may also affect their transcription. To investigate the correlation of transcription and promoter activity, we cloned 5' LTRs from 7 expressed proviruses into luciferase constructs and assessed their function following transfection of Tera-1 cells. The LTR Hs upstream of the 22q11.23 provirus was similarly assayed for activity to address its role in driving expression of the ancient provirus in lieu of the proviral 5' LTR 5B. The relative promoter activity for each assayed LTR was calculated as relative light units (RLU) normalized to that of a co-transfected control containing the SV40 promoter. Figure 4-8 shows RLU values reported relative to the most active promoter and FPKM values reported relative to the provirus with highest expression, as determined by the Plus stranded, Unique Only alignment detailed in Fig 4-3.

As shown in Fig 4-8, the 5' LTRs from most expressed proviruses displayed normalized promoter activities comparable to their associated FPKM. However, there is no direct interpretation for comparing FPKM and promoter activity values, as FPKM values represent relative transcript detection and RLU values represent the activity from the promoter associated with the transcript. In comparing the relative FPKM and relative RLU values, most provirus promoter activity varied less than 2.5-fold from the reported FPKM values. A major exception to this pattern was seen with the 22q11.23 provirus, whose LTR 5B was >500-fold less active relative to its FPKM, while the activity of the

upstream LTR Hs appeared to correlate with the FPKM of the 22q11.23 provirus. This result, taken together with the alignment data showing read-through transcription between the 22q11.23 LTR Hs and the downstream LTR 5B provirus (Fig 4-6), is consistent with the conclusion that the high expression of the 22q11.23 provirus in Tera-1 cells is due to the upstream LTR Hs, which is capable of high promoter activity in these cells. Another provirus whose promoter activity did not correspond well with its FPKM value was 3q13.2, which has predicted to be underrepresented in Fig 4-1. 3q13.2 displayed a 7-fold higher promoter activity level as compared to its FPKM, which may indicate that this provirus was underrepresented in the RNASeq analysis or that its position or regulation in the genome has a negative impact on 5' LTR promoter activity.

The canonical HML-2 LTR transcription start site is believed to be located at position 793 (Fig 4-9) [74, 154]. However, the RNASeq alignment showed 22q11.23 LTR Hs transcripts originating from further downstream, primarily at position 826 (Fig 4-6). Potentially, this LTR Hs can use an alternative start site to initiate transcription. To investigate this issue further, Tera-1 cells were transfected with a series of truncated LTR Hs constructs containing varying promoter elements (Fig 4-9) and analyzed for activity as described for Fig 4-8. We observed only small decreases in activity after truncating the 3' end of the LTR to position 805, an unexpected result, considering the RNASeq alignment data showing a TSS at 826 (Fig 4-6 and 4-9 B). No significant drop in activity was seen until the LTR was truncated to position 740, resulting in removal of both a GC box and the TATA box (Fig 4-9 A-B). In addition, truncations down to positions 522 and 460 ablated LTR activity almost entirely (Kruskal-Wallis, *$p<0.05$, **$p<0.01$, Fig 4-9 B).

**Figure 4-8. HML-2 promoter expression in Tera-1 cells.**

Comparison of the relative transcript expression level (FPKM; black) for a provirus and its corresponding relative luciferase expression level in Tera-1 cells transfected with a vector containing a luciferase reporter gene downstream of the indicated proviral 5' LTR. FPKM values are plotted relative to the highest expressed provirus, 22q11.23, which was set to 1 (left y-axis). RLU values are plotted relative to the most active HML-2 LTR, the 22q11.23 LTR Hs element, which was set to 1 (right y-axis). LTR activity is expressed as relative light units (RLU; gray) normalized to a control construct with a *Renilla* luciferase gene driven by an SV40 promoter. The relative promoter activities of the LTR Hs located 551bp upstream from the 22q11.23 provirus, the 5' LTR 5B of the 22q11.23 provirus and the 5' LTR Hs of six other expressed proviruses in Tera-1 cells are shown. The transfection experiments were performed by Meagan Montesion and supervised by Neeru Bhardwaj.

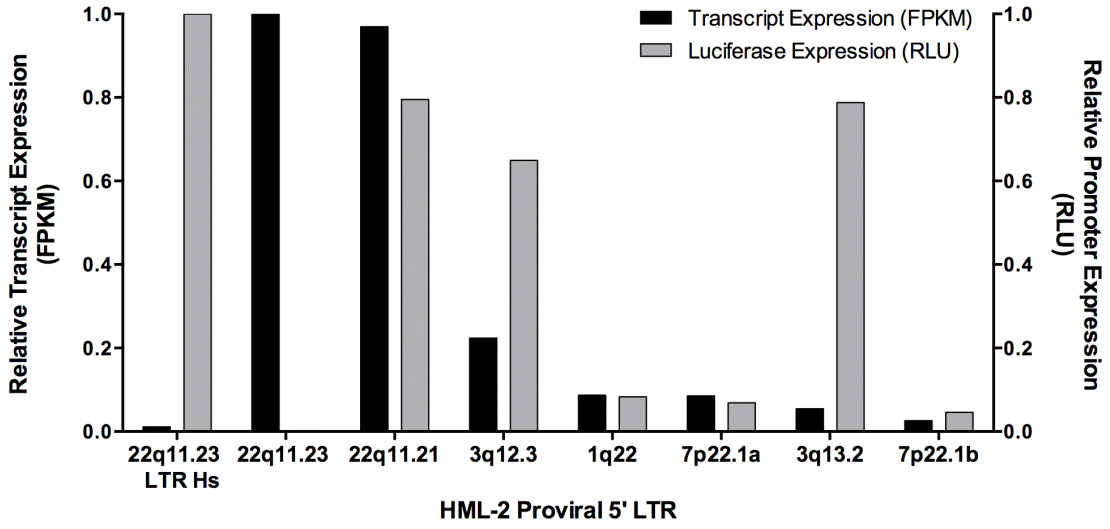
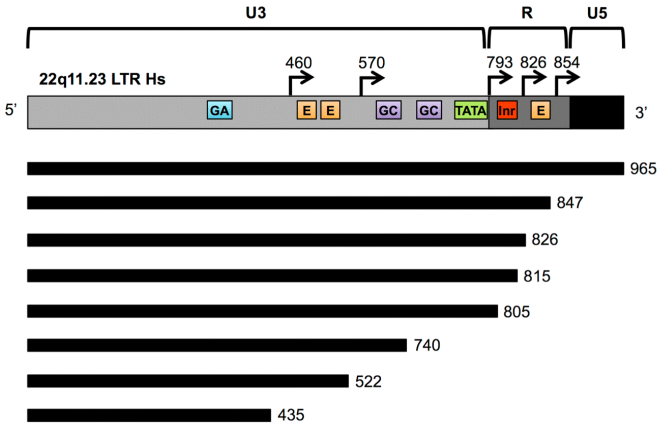

Transcript Expression (FPKM)
Luciferase Expression (RLU)
Relative Transcript Expression (FPKM)
Relative Promoter Expression (RLU)
1.0
0.8
0.6
0.4
0.2
0.0
22q11.23 LTR Hs
22q11.23
22q11.21
3q12.3
1q22
7p22.1a
3q13.2
7p22.1b
HML-2 Proviral 5' LTR

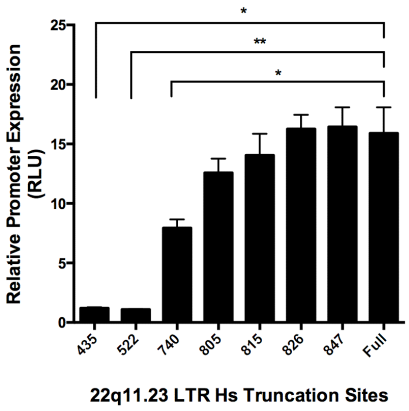
**Figure 4-9. The effect of truncations on 5' LTR promoter activity in Tera-1 cells.**

(A) Schematic of the 22q11.23 LTR Hs, showing the U3, R and U5 regions. Predicted transcriptional start sites are indicated with black arrows and nucleotide position. Colored boxes indicate previously described promoter element motifs [74, 125, 154]. Lines below the LTR diagram indicate the regions included in each truncated LTR construct, and numbers to the right of each line indicate the nucleotide position at which the LTR was truncated. GA, GA rich motif (nt 379-386, sequence GGGAAGGG); E, enhancer box (nt 465-476, sequence TTGCAGTTGAGA; nt 485-496, sequence AGGCATCTGTCT; nt 832-843, sequence CTCCATATGCTG); GC, GC rich motif nt 759-763, (sequence CCCCC; nt 602-606, sequence GGCGG); TATA, TATA box (nt 790-797, sequence AATAAATA); Inr, initiator element (nt 807-812, sequence CTCAGA). Cartoon is not drawn to scale. (B) Relative promoter expression levels of truncated 22q11.23 LTR Hs constructs in Tera-1 cells (Kruskal-Wallis, *$p<0.05$, **$p<0.01$). All luciferase experiments were conducted in triplicate and are shown as mean ± standard deviation. (C) Schematic of promoter motifs found in the 22q11.21 provirus 5' LTR Hs, the 22q11.23 LTR Hs and 22q11.23 provirus 5' LTR 5B. Crossed out boxes indicate presence of a mutation in the motif as compared to the canonical sequence. These experiments were performed by Meagan Montesion.

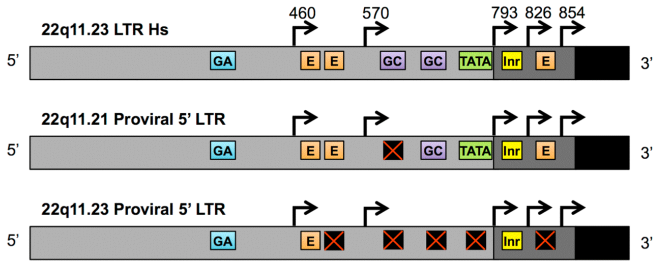
A


U3
R
U5
22q11.23 LTR Hs
460
570
793
826
854
5'
GA
E
E
GC
GC
TATA
Inr
E
3'
965
847
826
815
805
740
522
435

B


Relative Promoter Expression (RLU)
25
20
15
10
5
0
*
**
*
435
522
740
805
815
826
847
Full
22q11.23 LTR Hs Truncation Sites

C


22q11.23 LTR Hs
460
570
793
826
854
5'
GA
E
E
GC
GC
TATA
Inr
E
3'
22q11.21 Proviral 5' LTR
22q11.23 Proviral 5' LTR

Thus, the ability of the LTR Hs to promote transcription is largely dependent on GC box and/or TATA box promoter elements in the luciferase assay and did not appear to mimic the TSS exhibited in its native genomic location, although we have not yet directly determined its site of transcriptional initiation.

The most active LTRs in the Tera-1 cells, namely the 5' LTR Hs of the 22q11.21 provirus and the LTR Hs driving the 22q11.23 provirus, may have preserved promoter elements that allowed for their activity, in comparison to less active, older LTRs, like the 22q11.23 proviral 5' LTR 5B. Accordingly, canonical E box, GC box and TATA box elements were found in the most active LTRs (Fig 4-9 C). Conversely, the mostly inactive 22q11.23 LTR 5B displayed 1-3 nucleotide mutations for 5 out of the 8 promoter elements on its 5' LTR (Fig 4-9 C). Significantly, this LTR did not retain the canonical GC box or TATA box sequences, which were shown to be important for HML-2 LTR transcription in Fig 4-9 B. Thus, HML-2 LTR activity in Tera-1 cells appears to be in part reliant on maintenance of canonical promoter motifs.

*4.5 RNAseq HML-2 expression profiles of HIV-1 infected individuals*

The application of RNASeq to defining HML-2 expression at the proviral level was successful in both simulation and in implementation using the Tera-1 cell line [15]. Thus, this methodology was applied to the HIV-infected patient population in order to understand which HML-2 proviruses are overexpressed in HIV-1 infected individuals, a result originally quantified using the HML-2 *env* qPCR that detects HML-2 proviruses in bulk and cannot discriminate between differences in expression of individual loci (Fig 3-4A).

In the HIV-1 infected population, HML-2 expression appears to occur in multiple blood cell types and does not seem to be directly dependent upon HIV-1 replication, but rather an indirect consequence of infection (Fig 3-4 through 3-6) [14]. There are limited data on which proviruses are expressed during infection and whether the upregulated HML-2 proviruses carry open reading frames. Potentially, expression of retroviral proteins could have an effect on HIV-1 disease progression, or recombination between individually defective HML-2 proviruses could create an infectious variant, a process that occurs with endogenous retroviruses in other animals [230, 278].

In order to address these possibilities, HML-2 loci expressed during infection were determined by performing RNASeq on PBMCs from 10 HIV-1 infected and 6 uninfected individuals. These samples were taken from the HIV-1 population described in Table 3-1. The libraries were constructed from whole (unsorted) PBMC RNA in order to capture each patient's HML-2 expression profile. This strategy was undertaken since HML-2 provirus expression was not determined to specifically originate from one cell type; rather, it was detected in multiple cell types. Thus, for these samples, RNASeq was used to probe complex cell populations to determine proviruses that could have a pathogenic impact.

PBMC RNASeq libraries were prepared as described for Tera-1 cells with the exception that PBMC RNA was left unsheared prior to reverse transcription for the library preparation. By leaving the RNA unsheared, we maximized the size of the insert to be sequenced, potentially reducing the amount of multi-reads due to the longer sequence available for alignment. Alignment of reads to multiple closely related proviruses was observed in the analysis of HML-2 proviruses expressed in the Tera-1 cell

line. In practice, we achieved fragment sizes primarily ranging from 225-400 bases in length using this approach, greater than the ~190 base fragments of the Tera-1 cell library. RNASeq reads were trimmed to remove adaptor sequences and low quality bases. In addition, any trimmed reads with a length less than 100 bases were removed.

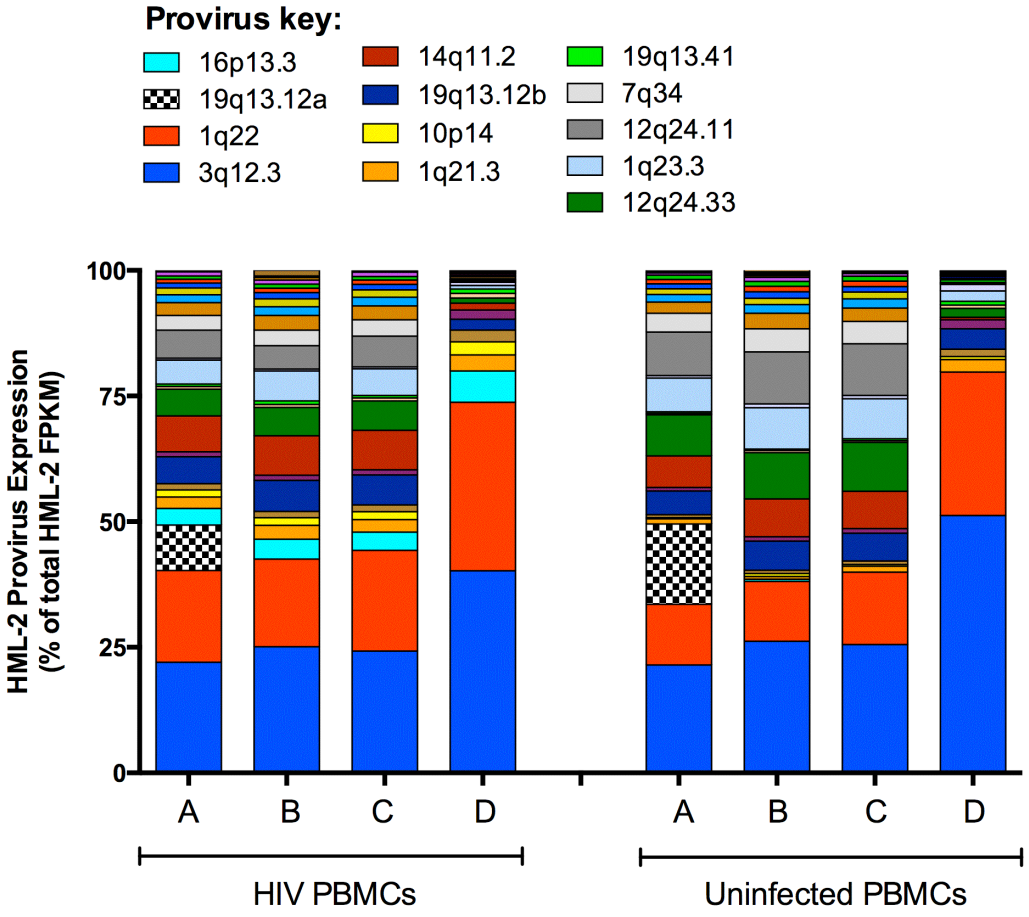
The RNASeq reads were aligned to the hg19 build of the human genome and the HML-2 reference genome, as with the Tera-1 alignments. These two sequence alignments were performed per subject for two main reasons: (1) in order to corroborate HML-2 provirus expression profiles; and most importantly, (2) to allow for the detection of elements that are not annotated in the available reference human genome. After alignment, reads were filtered to keep only uniquely aligned reads in order to offset any misrepresentation of HML-2 proviruses as being expressed. Expression analysis was performed using the program CuffDiff to determine differentially expressed genes between samples as well as CuffNorm to output normalized FPKM values for all samples tested, taking into consideration the strandedness of the read pairs.

In comparing the results of the two alignments, a discrepancy was discovered. All major hits were corroborated between the hg19 and HML-2 reference alignments except for the hg19 alignment detection of provirus 19q13.12a (shown in black/white checker print in Figure 4-10; compare columns A = hg19 alignment and B = HML-2 genome alignment). The read IDs for the reads aligning to the 19q13.12a genomic coordinates were extracted from the hg19 alignment file from the individual with the highest 19q13.12a provirus expression. When these read IDs were searched for in the subject's corresponding HML-2 genome alignment file, the reads were not found, thus confirming the lack of 19q13.12a detection in the HML-2 alignment and furthermore that these reads

were not aligned to a different provirus. When the extracted reads were then analyzed in BLAT to find their best genomic matches, a 100% match was found to be an RNA repeat (LSU-rRNA_Hsa) on an unassembled genomic contig (chrUn_gl000220). This type of repeat element is also inserted in the 19q13.12a provirus, at a 98.5% match. Thus, the detection of provirus 19q13.12a in the hg19 alignment was due to a misalignment of reads to a repeat element that happens to be present within the provirus. Though this repeat is present in the provirus, it was not included in the HML-2 reference genome since it is not of HML-2 origin. Since it does not represent true proviral expression, this proviral hit was excluded from analysis (Figure 4-10, column C). After this correction, the two alignments are now in agreement on expressed HML-2 loci.

**Figure 4-10. Summary of alignment methods and their effects on HML-2 detection.**

The proportion of individual HML-2 proviruses expressed in HIV-1 infected (n=10) and uninfected individuals (n=6) was summarized across biological replicates for each alignment condition. Average FPKM values calculated from uniquely aligned reads for each provirus were converted into a percentage value of total expressed HML-2 FPKM and plotted, as in Fig 4-3.




**Sample key:**

A = hg19 alignment
B = HML-2 genome alignment
C = hg19 alignment, corrected (no 19q13.12a)
D = hg19 alignment, corrected; Plus stranded

As shown in Figure 4-11 and 4-12 A, when considering the strandedness of reads, there is a marked drop off in HML-2 expression. In comparison, *GAPDH* expression was not affected by this analysis (Fig 4-12 B). In our previous analysis with Tera-1 cells, filtering for unique reads appeared to lower the number of reads aligned to HML-2 proviruses by ~47%, but removing antisense alignments affected only a minority expressed proviruses (Fig 4-2 A). Conversely, in the PBMC analysis, ~35% of the aligned HML-2 reads were in antisense orientation and filtering for uniquely aligned reads had little effect on the level of measured HML-2 expression (Fig 4-11). The lower amount of multi-reads reflects that many of the expressed proviruses in blood cells have distinct sequence. However, expression is not driven by the proviral 5' LTR that enables sense transcription, as seen in the increased level of antisense transcripts. In contrast, in Tera-1 cells, strong 5' LTR activity drives sense transcription of newer, highly similar integrations.

As shown in Figure 4-10 (compare columns C = Unstranded and D = Plus stranded), not all proviruses were affected equally in the stranded analysis. The highest expressed proviruses, 3q12.3 and 1q22, were not affected by the plus stranded alignment. However, less highly expressed proviruses present in introns of expressed genes (in parentheses) were reduced, including 19q13.12b (*ZNF420*), 14q11.2 (*DHRS4L1*), 12q24.33 (*ZNF140*), 1q23.3 (*CD48*) and 12q24.11 (*PPTC7*). These proviruses are inserted antisense to the direction of transcription for the gene in which they are located. In addition, the intronic sequences around the proviruses also show read alignment in the same transcriptional orientation as the gene. These observations are consistent with the explanation that transcript detection was a result of read-through transcription and this was detected due to

pre-mRNA being included in the total RNASeq library preparation. Interestingly, expression from provirus 1q23.3 was detected in a recent study using *env* amplicon sequencing to investigate HML-2 expression in lymphocytes [26]. The authors suggest that the defective 1q23.3 Env could pseudotype HIV particles and affect HIV pathogenesis, however our results show that 1q23.3 is transcribed primarily in antisense orientation and thus would not be translated into Env protein.

**Figure 4-11. Relative gene expression in PBMCs from HIV-1 infected and uninfected individuals.**

FPKM values for reference genes (*GAPDH*, *PPIA* and *HPRT1*), total HML-2 expression using different alignments (HML-2 Unstranded and Unfiltered, HML-2 Unstranded and Unique Only, and HML-2 Plus stranded and Unique only), the interferon regulated HIV restriction factor gene *APOBEC3G* and the immune marker gene *CD38* are shown for the HIV-1 infected (n=10) and uninfected (n=6) populations tested. Unless indicated, FPKM values are derived from plus stranded, unique alignments. Total HML-2 expression is calculated as a sum of all expressed HML-2 proviral FPKM. Error bars reflect the standard deviation between biological replicates. Statistical significance determined in CuffDiff after controlling for multiple comparisons, **p = 0.004 and *p = 0.03.

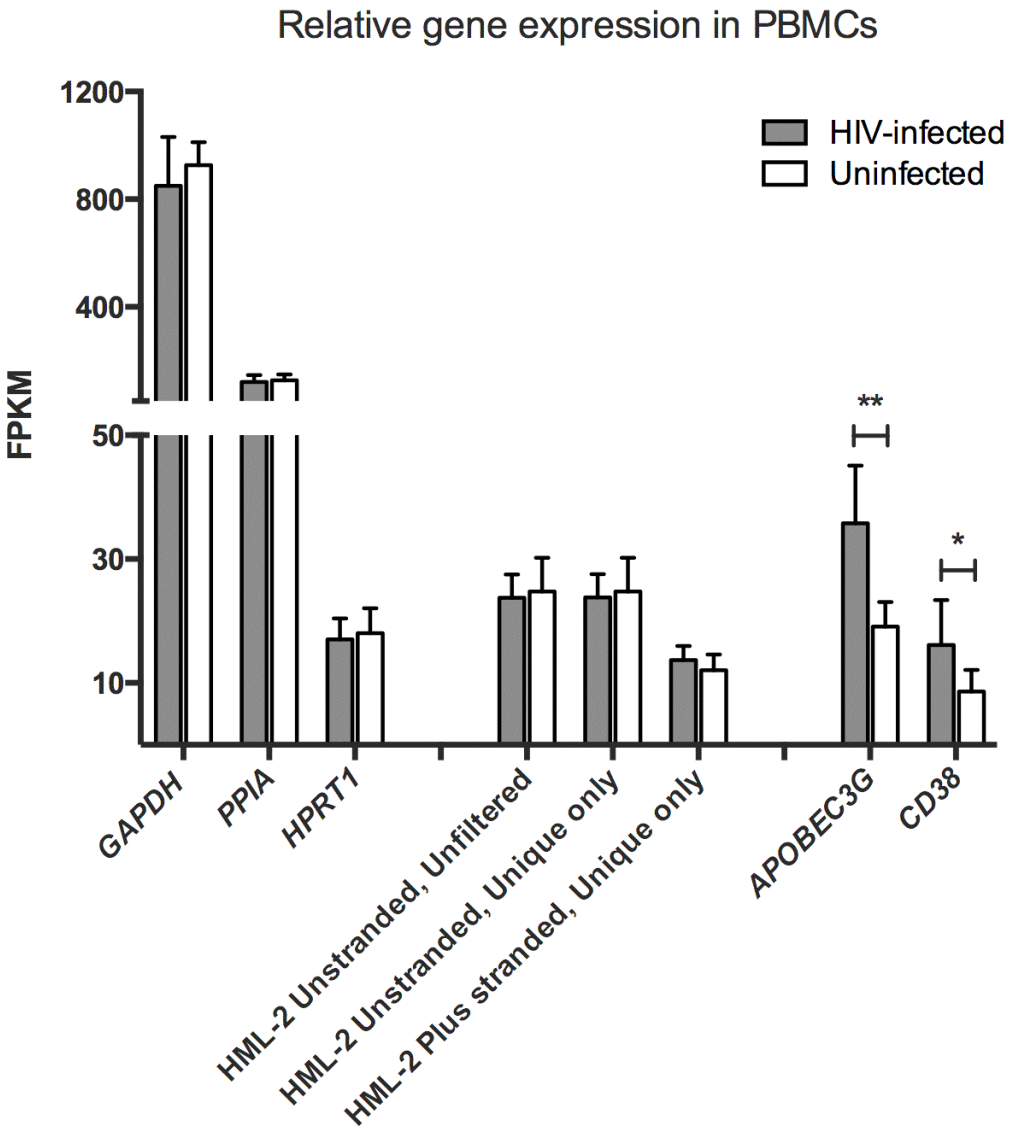
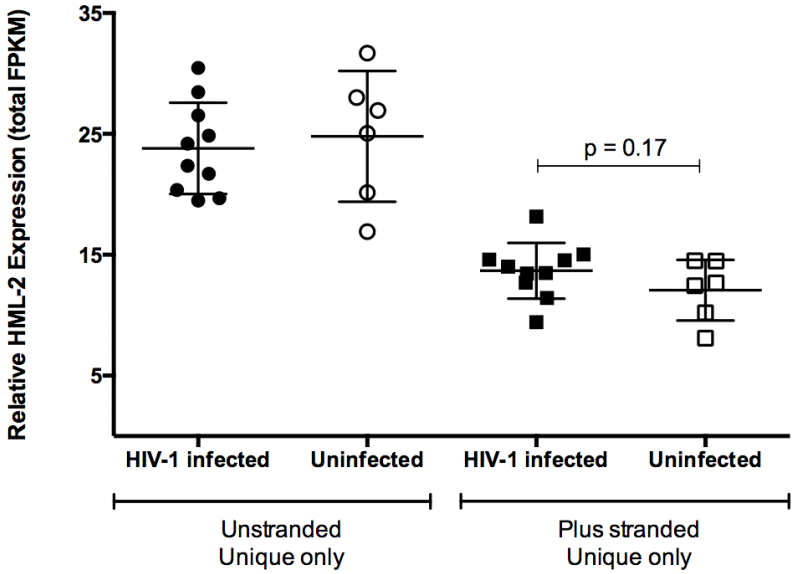

Relative gene expression in PBMCs
HIV-infected
Uninfected
FPKM
1200
800
400
50
30
10
**
*
GAPDH
PPIA
HPRT1
HML-2 Unstranded, Unfiltered
HML-2 Unstranded, Unique only
HML-2 Plus stranded, Unique only
APOBEC3G
CD38

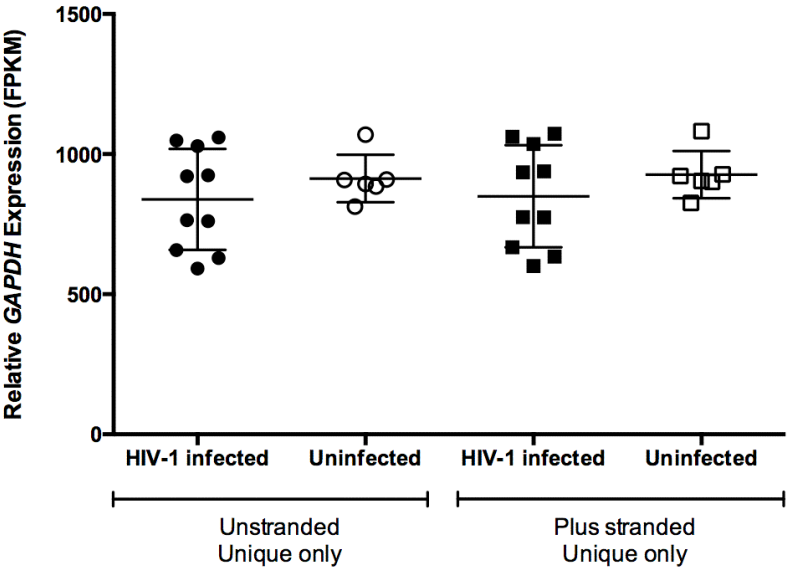
**Figure 4-12. Effect of eliminating antisense reads on relative gene expression.**

FPKM values for HIV-1 infected and uninfected individuals were plotted for each condition using data from hg19 human genome alignments. (A) Total HML-2 FPKM values were summed and plotted for each subject. (B) GAPDH FPKM values plotted for each subject. The p value was calculated using a Mann-Whitney *t*-test.

A


Relative HML-2 Expression (total FPKM)
35
25
15
5
p = 0.17
HIV-1 infected
Uninfected
HIV-1 infected
Uninfected
Unstranded
Unique only
Plus stranded
Unique only

B


Relative GAPDH Expression (FPKM)
1500
1000
500
0
HIV-1 infected
Uninfected
HIV-1 infected
Uninfected
Unstranded
Unique only
Plus stranded
Unique only

Another difference that arose between the hg19 and HML-2 reference genome alignments was that even though the same proviruses were being detected, a statistical significance in total HML-2 expression between HIV-1 and uninfected populations was only seen in the HML-2 genome alignment (Fig 4-13 A = HML-2 genome alignment, ***p = 0.0002; Fig 4-12 A for hg19, p = 0.17). A statistically significant difference in HML-2 expression was demonstrated using an *env*-specific qPCR previously (Fig 3-4 A). A potential reason for the greater sensitivity of the HML-2 genome alignment to pick up a significant difference between the populations is that only HML-2 reads were considered for comparison and library size normalization. As HML-2 expression is low compared to other cellular genes, the library size normalization in the hg19 alignment is driven by the mapping of reads to non-HML-2 genes and may obscure small differences. As a proof of principle, when only HML-2 annotations are considered for gene expression analysis and FPKM calculation in the hg19 alignment, a difference in expression between the HIV-1 infected and uninfected populations is now apparent (Fig 4-13 B; **p = 0.003). While a trend towards higher HML-2 expression in the HIV population was seen in the standard hg19 analysis (Fig 4-12 A, Plus stranded and Unique only), for low expressed loci like HML-2 proviruses, using default standardizations may impair data analysis and underestimate population differences.

Upon examining the individual proviruses expressed in the HIV-1 infected and uninfected conditions, it is apparent that the majority of expressed proviruses are shared and not unique to either condition (Fig 4-10, HIV – column D and Uninfected – column D). However, the magnitude of expression between populations can differ (Fig 4-14). The top two expressed proviruses in PBMCs in both populations are the LTR Hs

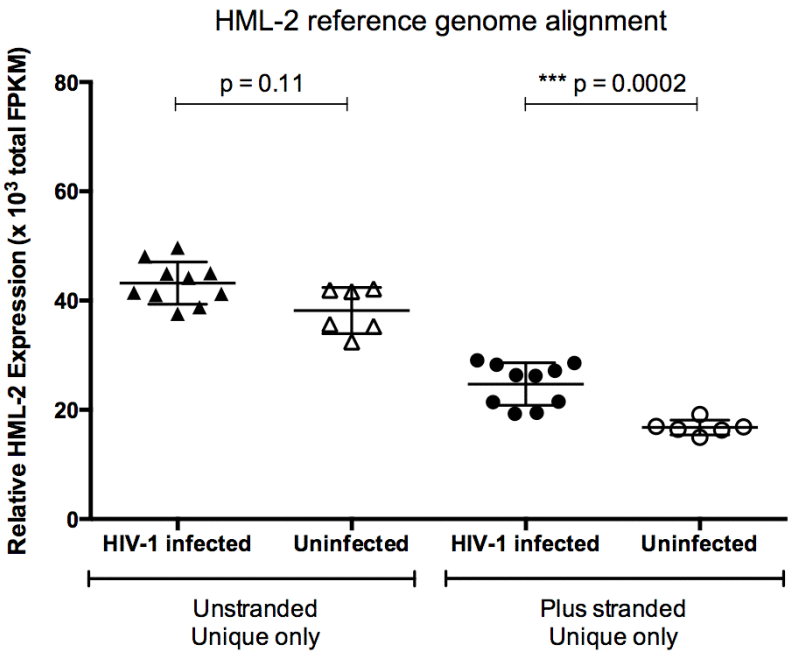
proviruses 3q12.3 and 1q22. Both proviruses trend towards higher expression in HIV-1 infected individuals, with 1q22 significantly higher, assuming a 10% false discovery rate. The expression of these proviruses are most likely driving the difference in total HML-2 expression between populations, as assessed by *env* qPCR and RNASeq. These two hits corroborate the results from a recent study investigating the expression of HML-2 *env* sequences in lymphocytes from HIV-negative patients [26]. 3q12.3 has an ORF for the structural gene *gag*, whereas 1q22 does not fully encode any of the essential retroviral genes *gag*, *pro*, *pol* or *env*. Notably, the truncated Env from 1q22 was able to be expressed and properly localize to the cell membrane in culture but could not pseudotype HIV or SIV virions [26].

Expression of proviruses at 3q12.3 and 1q22 could be enhanced due to their location near active transcription units and appears to be the result of transcription originating in upstream repetitive elements (Fig 4-15). 3q12.3 is located 5.2kb downstream of the highly expressed gene *RPL24* (HIV average FPKM = 1148, Uninfected average FPKM = 1789). Though transcription is in sense orientation, it does not appear that the 3q12.3 5' LTR is driving its own transcription. In the alignment shown in Fig 4-15, transcription originates in an upstream repetitive element, specifically a type of SINE element called AluSx3 (chr3:101407448-101407756). Transcription continues downstream to include the 3q12.3 proviral sequence and primarily terminates in the 3q12.3 3' LTR. The second highest expressed provirus, 1q22, is located between the genes *MSTO1* (11.7kb upstream; HIV average FPKM = 3.2, Uninfected average FPKM = 2.9) and *YY1AP1* (23.6kb downstream; HIV average FPKM = 24.8, Uninfected average FPKM = 28.9). Interestingly, many of the reads aligning to the 1q22 5' LTR are actually spliced reads

**Figure 4-13. HML-2 provirus expression is upregulated in HIV-1 infected individuals.**

Total FPKM values for HML-2 proviruses expressed in HIV-1 infected and uninfected individuals were summed and plotted for each condition. (A) Total HML-2 FPKM values were summed and plotted for each patient using data from the HML-2 genome alignment. (B) Total HML-2 FPKM values were summed and plotted for each patient using data from the hg19 alignment, with only HML-2 annotations analyzed for gene expression and used for library size normalization. The p values were calculated using Mann-Whitney *t*-tests.

A


HML-2 reference genome alignment
Relative HML-2 Expression (x 10^3 total FPKM)
80
60
40
20
0
p = 0.11
*** p = 0.0002
HIV-1 infected
Uninfected
HIV-1 infected
Uninfected
Unstranded
Unique only
Plus stranded
Unique only

B

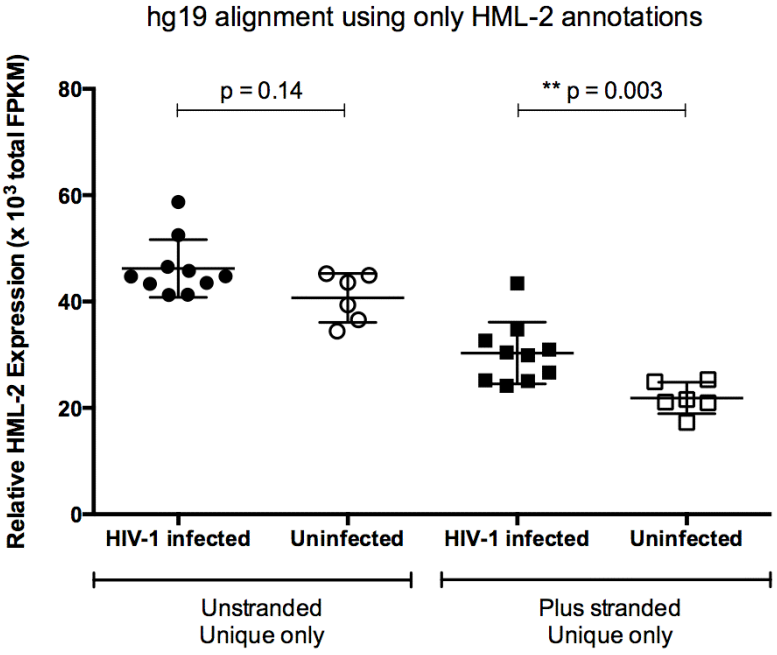

hg19 alignment using only HML-2 annotations
Relative HML-2 Expression (x 10^3 total FPKM)
80
60
40
20
0
p = 0.14
** p = 0.003
HIV-1 infected
Uninfected
HIV-1 infected
Uninfected
Unstranded
Unique only
Plus stranded
Unique only

**Figure 4-14. The most highly expressed HML-2 proviruses in PBMCs.**

FPKM values for the top expressed HML-2 proviruses in HIV-1 infected and uninfected individuals were calculated based on the Plus stranded, Unique only alignment using the hg19 reference (as in Fig 4-13 B). The p values were calculated using Student's *t*-tests with the Holm-Sidak post-test to correct for multiple testing. The p values remained significant assuming a 10% false discovery rate.

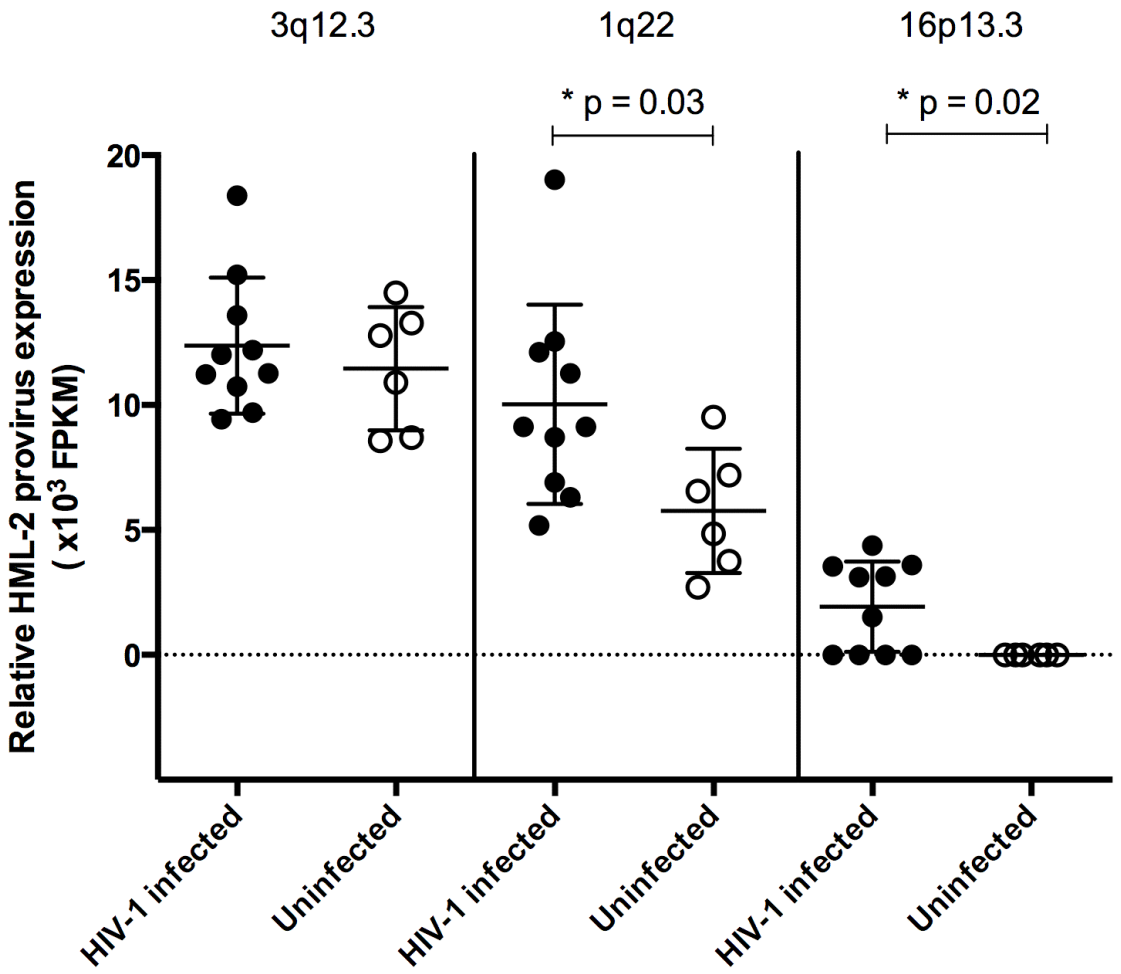
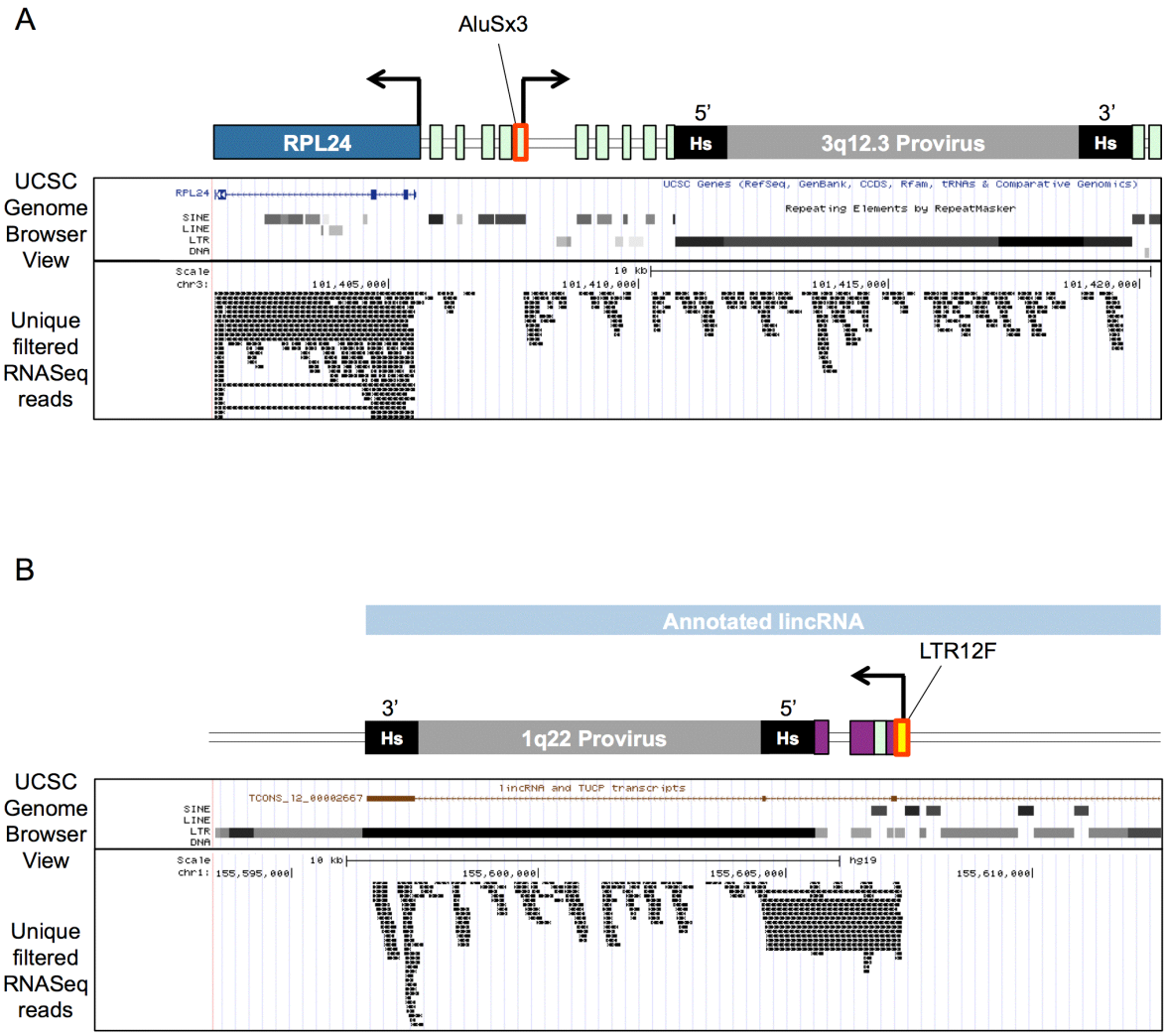

3q12.3
1q22
16p13.3
* p = 0.03
* p = 0.02
Relative HML-2 provirus expression
( x10^3 FPKM)
20
15
10
5
0
HIV-1 infected
Uninfected
HIV-1 infected
Uninfected
HIV-1 infected
Uninfected

**Figure 4-15. UCSC genome browser view of reads aligned to 3q12.3 and 1q22.**

RNASeq reads generated from the PBMCs of an HIV-1 infected patient were aligned to the hg19 build of the human genome, filtered for Unique Only reads and a BEDfile of the alignment was uploaded to the UCSC genome browser as a custom track. Screenshots of the UCSC genome browser around the LTR Hs proviruses (A) 3q12.3 (chr3:101410737-101419859); and (B) 1q22 (chr1:155596457-155605636) are shown. A cartoon is placed above the UCSC Genome browser views to clearly indicate the inferred start of transcription for each provirus, namely in an AluSx3 element for 3q12.3 and an ERV LTR12F for 1q22. The black arrow indicates the start site and direction of transcription observed in PBMCs and a red outline surrounds the element where transcription originates. HML-2 proviral LTR sequence is shown in black and the internal genomic sequence (*gag-pro-pol-env*) is shown in gray. The UCSC Genome browser view shows the annotations for repetitive elements, known transcripts (*RPL24*) and lincRNA (lincRNA accession number: TCONS_l2_00002667, chr1:155596547-155618335). Of note, short interspersed nuclear elements (SINEs), like AluSx3, are shown in light green and HERV-W sequence is shown in purple. A key depicting the symbols used for reads crossing splice junctions and reads that do not is shown below the screenshots.




Reads crossing splice junctions
(Note: appear as one line of uniform boxes spanning 100s-1000s nt)

Reads that do not cross splice junctions
(Note: appear as a grouping of 1-3 boxes and span 100-301 nt)

originating in an unrelated proviral LTR (ERV1, LTR12F; chr1:155607243-155607452), as shown in Fig 4-15. This transcript of provirus 1q22 partially overlaps with an annotated lincRNA found to be highly expressed in white blood cells and breast tissue (TCONS_l2_00002667; chr1:155596547-155618335).

Compared to these two proviruses, the FPKM values for the other expressed proviruses were much lower (Fig 4-16). All of the other detected proviruses are ancient integrations, with estimated ages of over 2 million years, except for 8p23.1a which is a human specific integration retaining all ORFs. Of interest, the LTR5B provirus 16p13.3 was only detected in HIV-infected population in the hg19 alignment. It is located in an intron of the gene *FLYWCH1*, which was expressed at low but equal levels in HIV-1 infected and uninfected individuals (HIV average FPKM = 3.3, Uninfected average FPKM = 3.5). However, it does not have any complete open reading frames and the significance of its expression is unclear. It appears to be the outcome of read-through transcription in an intron as there is not complete genome coverage.

In this study, the HIV-1 infected patients analyzed were all off therapy and exhibited a wide range of plasma viral loads (median: 3.74 $\log_{10}$ copies/mL, range: 1.95-5.55 log10 copies/mL). HIV replication was not correlated with expression of 3q12.3 (p=0.41), 1q22 (p=0.21) or 16p13.3 (p=0.52). However, it was associated with the expression of the activation marker *CD38* in this patient group (Fig 4-11; ***p=0.0008, Spearman r = 0.90), known to be expressed on highly activated CD4+ and CD8+ T cells from HIV-1 infected individuals and serving as a positive control in this analysis [178].

The two most highly expressed proviruses, 3q12.3 and 1q22, are both fixed in the human genome and are LTR Hs proviruses estimated to have integrated either 5.5-10 or

<2 million years ago respectively [233]. As mentioned previously, 3q12.3, which is not differentially expressed, retains a *gag* ORF but 1q22 does not encode intact essential genes. However, the provirus 1q22 is a type 1 HML-2 provirus capable of encoding the accessory protein Np9, which is the product of a double spliced transcript that may have oncogenic effects in a cell [98]. In our alignments, spliced reads indicating expression of *np9* were not observed. Potentially, use of deeper sequencing on these libraries or targeted qPCR for detection of 1q22 *np9* transcripts would be able to capture whether these are expressed in the HIV-1 population, and furthermore, if there is any association with a certain cell type or HIV disease characteristic.

There does not appear to be high potential for complementation and recombination to occur between expressed HML-2 proviruses. Though 3q12.3 encodes full-length Gag protein, it was not differentially expressed between populations and virions were not detectable in the plasma of HIV-1 infected patients. The human specific LTR Hs provirus 8p23.1a, which integrated at least 1.1 million years ago [110] and carries ORFs for *gag*, *pro*, *pol* and *env*, was only very weakly detected in this analysis. However, it is not infectious as-is and it is known that its encoded Env is non-fusogenic and cannot pseudotype SIV [59]. The other proviruses with ORFs shown in Fig 4-16 are from older integrations and have unknown functionality. These results are in stark contrast to our study on Tera-1 cells, where the two most highly expressed proviruses both carried full or nearly-full open reading frames for *gag* and produced viral like particles easily detectable in the cell supernatant. In Figure 4-17, a comparison of HML-2 expression between PBMCs and the teratocarcinoma cell line Tera-1 is shown. While 3 out of the top 4 HML-2 proviruses expressed in PBMCs are transcribed in Tera-1 cells, the HML-2

profile of Tera-1 cells is vastly different, where these proviruses make up only a small proportion of what is expressed.

**Figure 4-16. Heatmap of HML-2 provirus expression across individuals.**

RNASeq reads derived from HIV-1 infected and uninfected PBMC RNA were aligned to the hg19 build of the human genome, using the Plus stranded, Unique only alignment. FPKM values representing relative expression were determined for all expressed HML-2 proviruses and log-normalized for use in heatmap generation. Log-normalized FPKM is shown from high (red) to low (blue) expression, as indicated in the key to the left. Intact ORFs are displayed for each provirus on the right with red crosses.

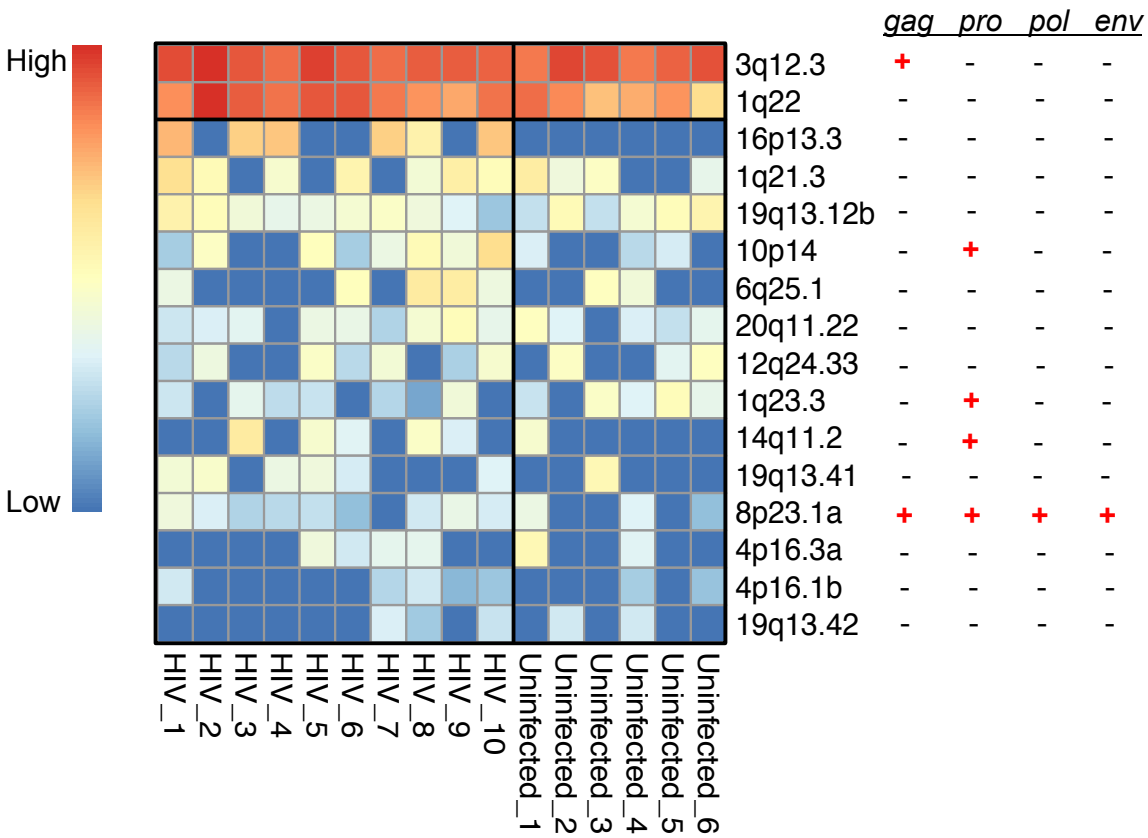
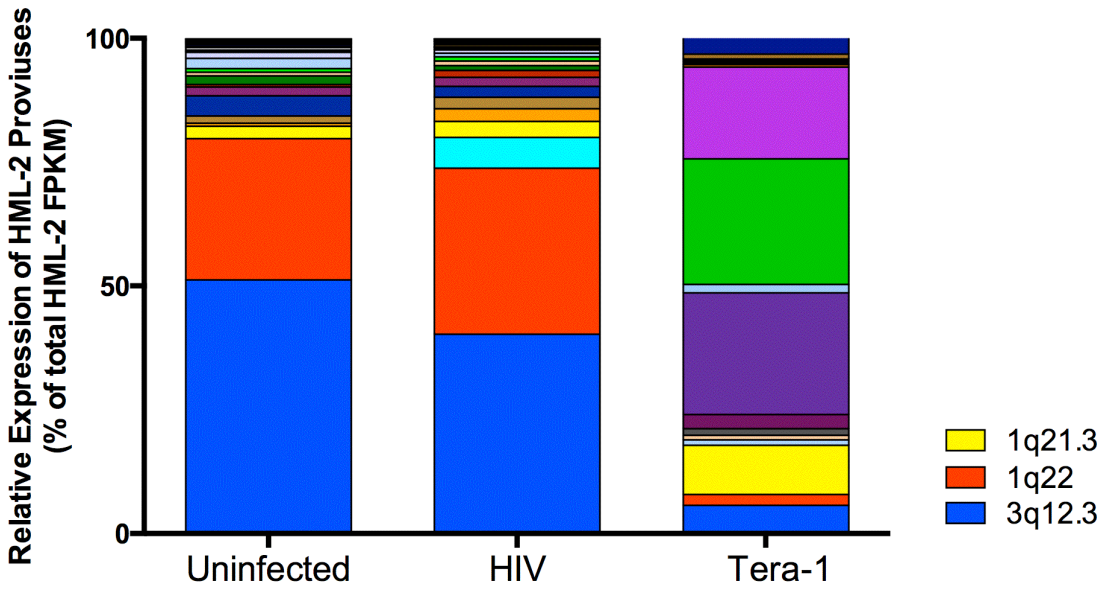

High
Low
HIV_1
HIV_2
HIV_3
HIV_4
HIV_5
HIV_6
HIV_7
HIV_8
HIV_9
HIV_10
Uninfected_1
Uninfected_2
Uninfected_3
Uninfected_4
Uninfected_5
Uninfected_6
gag
pro
pol
env
3q12.3 + - - -
1q22 - - - -
16p13.3 - - - -
1q21.3 - - - -
19q13.12b - - - -
10p14 - + - -
6q25.1 - - - -
20q11.22 - - - -
12q24.33 - - - -
1q23.3 - + - -
14q11.2 - + - -
19q13.41 - - - -
8p23.1a + + + +
4p16.3a - - - -
4p16.1b - - - -
19q13.42 - - - -

**Figure 4-17. Comparison of HML-2 proviral expression patterns in PBMCs and the teratocarcinoma cell line Tera-1.**

The proportion of individual HML-2 proviruses expressed in PBMCs from HIV-negative patients, PBMCs from HIV-1 infected patients and Tera-1 cells were summarized for each cell type. The average FPKM for each provirus was converted into a percentage and plotted according to (average provirus FPKM)/(average total HML-2 FPKM) x 100. The three proviruses shown in the figure key are highly expressed in PBMCs, but are at a lower prevalence in Tera-1 cells, a cell line known to produce HML-2 particles.


Relative Expression of HML-2 Proviuses
(% of total HML-2 FPKM)
100
50
0
Uninfected
HIV
Tera-1
1q21.3
1q22
3q12.3

## Chapter 5: Discussion

Excerpts from this chapter were previously published in:

Bhardwaj N, Maldarelli F, Mellors J, Coffin JM. "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. 2014;88(19):11108-20.

Bhardwaj N, Montesion M, Roy F, Coffin JM. "Differential expression of HERV-K (HML-2) proviruses in cells and virions of the teratocarcinoma cell line Tera-1." Viruses. 2015;7(3):939-68.

*HML-2 expression in HIV-1 infected plasma and cells*

HERV-K (HML-2) proviruses represent the most recently integrated proviruses in the human genome, some of which maintain ORFs for retroviral genes [233]. Although not expressed highly in most normal tissue, multiple studies have shown HML-2 transcription to be associated with several disease states, notably in cancers and HIV-1 infection [98, 203, 256]. Previous reports investigating HML-2 activity during HIV-1 infection have described detectable HML-2 virions in patient plasma, often at levels higher than HIV [42-44, 46, 47], HML-2 RNA upregulation in patient CD4+ and CD8+ T cells [179], and the presence of cytotoxic and humoral immune responses specific to HML-2 antigens [81, 164, 165, 222, 242]. The observations of HML-2 RNA and protein expression during HIV-1 infection have led to a proposal for development of a vaccination strategy to eliminate HIV-1 infected cells, which may be targeted by their ability to display HML-2 antigens [112, 204, 224]. However, the mechanisms governing HML-2 expression are unclear and, given the large number of proviruses, likely to be

very complex, and the cells expressing HML-2 RNA and protein are largely undefined. An impediment to investigating their possible role in disease has been the incomplete understanding of their expression patterns in human cells and, along the same lines, the availability of only small sample sizes of analyzed tissues [72, 221]. In addition, the exact proviral loci and the details of their regulation that may be important in a disease context are not known.

In this study, we investigated HML-2 expression in a cohort of HIV-1 infected patients to determine the cell source of HML-2 expression and the relationship of HML-2 expression to HIV disease status. Using a qPCR assay capable of detecting RNA from more than half of the known HML-2 elements (Fig 3-1), we observed, contrary to previous reports, that HML-2 virions were not detectable in patient plasma during untreated HIV-1 infection (Fig 3-2). The discrepancy between this and previous results was not due to differences in methodology, since our finding was confirmed using a previously published extraction and detection method with a subset of the clinical samples (data not shown). Similarly, a recent study using deep-sequencing techniques did not uncover an increase in overall HERV RNA or HERV-K RNA (inclusive of HML-2) sequences from the plasma of HIV-1 subtype B subjects as compared to uninfected controls [138]. It is possible that virion production occurs below our limit of detection using the *env* qPCR, as we did not achieve single copy sensitivity, but virion production would still be orders of magnitude less than reported.

During this analysis, we found that HIV-1 infected patient plasma showed a significant increase in extracellular DNA as compared to that from uninfected individuals (Fig 3-2). Plasma DNA, specifically mitochondrial DNA, has been observed in HIV-1

infected patients previously, though not at a level significantly different from controls [130]. The finding of extracellular DNA in the plasma of HIV-1 patients could help explain the discrepancies in our results from previous publications. If extracted plasma RNA was not treated with DNase and adequate reverse transcriptase negative controls were not run, our assay could have led to the erroneous detection of HML-2 RNA in patient plasma (Fig 3-2). However, the signal would have derived solely from extracellular DNA present in the plasma of HIV-1 patients and not from HML-2 RNA. This artifact is possible because HML-2 proviruses are present at high copy number in the human genome, with more than 90 proviruses per haploid genome [233]. Furthermore, levels of DNA in patient plasma could also be affected by sample handling: if blood samples were not immediately processed, increasing amounts of DNA could be present in patient plasma samples and provide more template for erroneous HML-2 RNA detection. Finally, quantitation of contaminating genomic DNA with an RNA-based standard could further inflate HML-2 signal due to the less than 100% efficiency of the RT step used for the standard curve. Therefore, a combination of the above factors could lead to the false detection of high levels of HML-2 RNA and possibly explain why previous publications have reported detection of HML-2 RNA in HIV-1 patient plasma.

The same group that reported the presence of HML-2 virions in the plasma of HIV-1 patients has recently explored the idea that HML-2, unlike any other betaretrovirus, may replicate using a packaged viral DNA genome instead of RNA [64]. This type of replication cycle is known to occur in the spumavirinae sub-family of *Retroviridae*, which is distantly related to betaretroviruses. Thus, our result of contaminating DNA in the plasma of HIV-1 patients harboring HML-2 signal could be explained by this

alternative hypothesis where virions package HML-2 DNA. Specifically, in the teratocarcinoma cell lines Tera-1 and NCCIT, in lymphoma patients and with a reconstituted HML-2 virus, HML-2 RNA was reported to undergo reverse transcription in the cell prior to packaging in a virion, or while in the virion and prior to infection of a new cell [64]. Furthermore, ~40% of all genomes packaged into virions were reported to be DNA [64]. However, in their previous finding of HML-2 virions in HIV-1 infected patients, these same authors did not note the presence of HML-2 DNA in the RT-controls but rather reported them as free of signal [46], which is a direct contradiction to their current stance. The relationship between these two reports by the same group remains to be clarified and an independent lab should critically test the presence of packaged HML-2 viral DNA in virions. If a betaretrovirus like HML-2 were to replicate akin to a spumaretrovirus, it would likely represent an independent evolution of a complex lifecycle known only to occur in a retrovirus with a drastically different genome organization.

Extracellular dsDNA could potentially function as a damage-associated molecular pattern if taken up by cells and detected by a dsDNA receptor in the endosome or cytosol, leading to immune activation [185]. Signaling due to extracellular DNA could potentially represent an additional pathway to chronic immune activation commonly seen in HIV-1 infected individuals [51]. In this study, association of plasma DNA with immune activation markers was not assessed through flow cytometry, though the immune activation marker CD38, present on activated T cells in HIV-1 infected individuals [178], was upregulated in the same HIV-1 population as assessed by RNASeq (Fig 4-11). Interestingly, the RNASeq screen showed that the cytosolic dsDNA sensors *AIM2*, *IFI16*

and *DDX60* were upregulated in the HIV-1 population over the uninfected population (data not shown; p values corrected for multiple testing are p=0.001, p=0.03 and p=0.001 respectively). Detection of dsDNA by AIM2 triggers inflammasome assembly, which is a protein complex that leads to post-translational processing of the pro-inflammatory pro-form cytokines IL-1β and IL-18 and induces pyroptosis, an inflammatory type of cell death [185]. IFI16, which is an AIM2-like receptor, is thought to either activate the inflammasome or lead to type 1 interferon upregulation through its downstream interactions [185]. Previously, IFI16 was associated with pyroptosis of CD4+ T cells with unproductive HIV-1 infection in the lymphoid compartment [167]. Both IFI16 and AIM2 were shown to bind DNA in a sequence-independent manner, potentially allowing for sensing of host DNA in the cytoplasm by binding the sugar-phosphate backbone [180]. DDX60 appears to stimulate type 1 interferon signaling after DNA recognition and is also capable of recognizing viral RNA in concert with the RNA sensor RIG-I [185]. There is the potential for these DNA sensing pathways to lead to increased immune activation in cells; additionally, since sensing can also lead to cell death by pyroptosis, their activation could contribute to the expulsion of DNA into the extracellular compartment and drive DNA signaling in a positive feedback loop. The relative contribution of extracellular DNA uptake to sensor activation is not apparent since HIV replication alone can lead to engagement of a DNA sensing pathway, as seen with IFI16 [167], though HIV replication was not shown to strongly activate AIM2 signaling [35]. In our correlation analysis, we found no significant correlation of HIV-1 replication, as measured by plasma viral loads, with the relative transcript levels of any of the aforementioned DNA sensors (data not shown). An additional study examining the levels

of plasma DNA in patients on successful antiretroviral therapy could be informative, as chronic activation is present, albeit at a lower level, in patients on long-term therapy and is associated with decreased longevity, cardiovascular disease and metabolic syndrome [189].

A significant upregulation in HML-2 RNA was assessed in PBMCs from HIV-1 infected patients in the absence or presence of antiretroviral therapy (***$p<0.0001$, Fig 3-4; *$p=0.02$, Fig 3-5). On average, the level of expression was about 2-fold higher than that seen in uninfected controls, although some patients showed levels of HML-2 transcription equivalent to HIV-negative subjects. This result can mean that HML-2 upregulation is not a universal phenomenon, or perhaps that the cell source of expression was not in high abundance in the patients with no measurable increase in transcription. The initial analysis was performed using unsorted PBMCs (Fig 3-4), so cell subset frequency could have had affected overall measured expression levels. To assess this possibility, total PBMCs from HIV-infected and uninfected patients were sorted into CD4+ and CD8+ T cells subsets, in addition to B cells and monocytes. These cell types represent the main constituents of PBMCs, although smaller populations including dendritic cells and NK cells could potentially represent sources of HML-2 expression. When HML-2 RNA upregulation was assessed in sorted cells, we saw no significant difference in expression in any subset from HIV patients as compared to controls (Fig 3-6), although each subset individually showed a small increase in patients compared to controls, with the greatest difference in monocytes, a cell type that is not a target for HIV infection in vivo [115]. The lack of enrichment in the CD4+ T cell population is consistent with HML-2 RNA expression having no clear relationship to HIV replication

(Fig 3-5), while the lack of enrichment in other cell populations could exemplify the indirect mechanism regulating HML-2 expression in HIV-1 infected patients. The differences in the magnitude of HML-2 transcription in cell subsets could be due to cell-specific transcription factors, methylation patterns or other epigenetic changes, which are believed to control endogenous retrovirus expression in differentiated tissues [150, 154]. Based on these results, it appears that differential expression from multiple cell sources may lead to the 2-fold difference in HML-2 expression in HIV-infected versus control individuals.

It is important to point out that our results do not exclude the possibility of significant upregulation of expression of HML-2 in HIV-infected cells. On average, less than 0.2% of CD4+ T cells are infected with HIV during chronic infection [114], so even a 10-fold upregulation of HML-2 RNA would only lead to a 2% increase in the total cell population. To answer this question directly, it will be necessary to sort either HIV- or HML-2-expressing cells from patient samples, which is a daunting task, or alternatively address HML-2 upregulation in *ex vivo* HIV infection of primary CD4+ T cells.

In this study, no significant correlation was observed between HIV-1 disease markers, including level of viremia, intracellular HIV RNA and DNA, and HML-2 expression in patient PBMCs (Fig 3-4 and 3-5) and no enrichment was seen in HIV-1 target CD4+ T cells, as previously mentioned (Fig 3-6). In addition, there was no effect of antiretroviral therapy on HML-2 expression (Fig 3-5). The lack of effect of HIV replication or antiretrovirals is interesting as they have been reported to have positive and negative effects respectively on HML-2 virion production in HIV-1 infected patients [43, 44]. *In vitro*, the accessory proteins Tat and Vif have been reported to positively influence HML-

2 RNA and protein expression [48, 86, 112]. The difference between the *in vitro* and *in vivo* results may be due to the higher levels of infection and therefore protein production seen during *in vitro* infections, or potentially a non-canonical function the accessory proteins assume under *in vitro* conditions.

*HML-2 profiling of HIV-1 infected patient cells*

As the *env* qPCR measures total HML-2 signal and does not distinguish among the proviruses expressed, it is possible that individuals expressed distinct proviruses or that the expression of only some HML-2 proviruses was associated with HIV disease. To address these issues, we applied RNASeq analysis to capture the HML-2 transcription profile of HIV-1 infected patients. Contrary to our expectations, we found that the identities of expressed HML-2 proviruses in PBMCs were remarkably consistent between HIV-1 infected and uninfected individuals (Fig 4-10, 4-16, 4-17).

The highest expressed proviruses in each case were the LTR Hs proviruses 3q12.3 and 1q22, followed by a combination of lesser-expressed elements. Both 3q12.3 and 1q22 are fixed in the genome, thus are present in all tested individuals [233]. This finding corroborates a previous report showing expression of 3q12.3 and 1q22 in uninfected peripheral blood lymphocytes [26] and extends it to the HIV-1 infected population. In our study, we found that 1q22 was expressed significantly higher in HIV-1 infected individuals, assuming a 10% false discovery rate (*p=0.03; Fig 4-14). Interestingly, the T cell line model KE37.1-IIIB, which is chronically infected with HIV-1, showed higher expression of the provirus 1q22 as compared to the uninfected cell line KE37.1; however, it did not recapitulate expression of 3q12.3 [259]. This observation may imply that this

cell line model cannot capture the complexity of HML-2 expression, or that T cells only express a restricted set of HML-2 proviruses. However, our study and a previous study [26] analyzed mixed cell populations so this difference remains untested, though it seems likely that cell line expression differs from that of primary cells.

The fixed HML-2 LTR 5B provirus 16p13.3, which does not carry the canonical retroviral ORFs, was only detected in the HIV-1 infected group using the hg19 alignment. However, 1 uninfected individual out of the 6 tested was detected as having 16p13.3 expression in their corresponding alignment to the HML-2 reference genome (data not shown). Expression of 16p13.3 did not appear to be related to the expression level of gene *FLYWCH1*, which contains the proviral insertion in an intron, as gene expression was equivalent between populations (Fig 4-14). The enriched expression of this provirus in HIV-1 infection was unique, as this pattern was not observed for any other HML-2 proviruses detected. However, 16p13.3 did not show full genome coverage in alignment, a finding that weakens the reliability of this measurement.

The only polymorphic provirus detected in the hg19 alignment was 8p23.1a, which is capable of encoding all retroviral gene products but was detected only very weakly in this analysis (Fig 4-16). The frequency of this provirus detected in our sample set was much higher than anticipated, where 9/10 HIV-1 infected individuals and 3/6 uninfected individuals exhibited expression of this provirus (Fig 4-16). However, there was not full genome coverage and signal was limited to a few reads in *gag* in the alignments. In the population, the 8p23.1a provirus is insertionally polymorphic and is expected to only be present at a frequency of 35% in African-Americans, 39% in Hispanics and 6% in Caucasians [110]. Based on ethnicities provided for our HIV-1 population, 100% of

African-Americans (2/2), 75% of Hispanics (3/4) and 100% of Caucasians (4/4) carry this provirus. Due to the very low detection of this provirus and insertion frequencies inconsistent with published observations from large datasets, it appears that 8p23.1a detection may be incorrect and a result of misalignment, potentially from the related provirus 1q22 (Fig 4-1). The actual carriage of this provirus remains to be validated using PCR analyses targeting the integration site in PBMCs from our HIV-1 infected and uninfected populations. However, the likely inaccurate detection of 8p23.1a is consistent with the fact that 3q12.3 and 1q22, the highest expressed proviruses, represent the functional limit of provirus identification, as the other proviruses are too poorly expressed and alignments too sparse to consider them definitively transcribed.

Our study did not confirm the finding that the provirus 1q23.3, which encodes a defective Env protein, is highly expressed in lymphocytes [26]. In fact, the 1q23.3 provirus was only detected as transcribed in the antisense orientation (Fig 4-10), most likely due to its presence in the intron of *CD48*, which is transcribed in the opposite orientation as the provirus. This result contradicts the hypothesis that Env protein generated from 1q23.3 transcription could be used to pseudotype HIV-1 and interfere with replication, as was proposed [26]. Additionally, in the report cataloguing HML-2 expression from the HIV-1 infected cell line KE37.1-IIIB, the authors identified 7q34 expression after PCR and sequencing of HML-2 amplicons [259]; however, in our study, this transcript was identified as a product of antisense transcription in lymphocytes as well (Fig 4-10). In addition to 1q23.3 and 7q34, many proviruses were expressed in this manner, where antisense proviral transcription was detected due to read-through transcription or their presence in introns of expressed genes (Fig 4-10, 4-11). When only

considering reads aligning to HML-2 proviruses in the sense orientation, the level of total HML-2 expression was reduced by ~35% (Fig 4-11). The orientation of transcription should be considered in order to accurately interpret the possible consequences of HML-2 expression in a cell.

Though they are transcribed in sense orientation, the theme of HML-2 provirus expression from areas of transcriptional activity held true with 3q12.3, 1q22 and 16p13.3, as each of these proviruses is located near or in genes expressed in both HIV-1 infected and uninfected individuals (Fig 4-15, 5-1). It appears likely that the active transcriptional environment supports the expression of these proviruses in PBMCs. Furthermore, transcription does not appear to originate in the 5' LTR, as would be anticipated in provirus driven transcription. Rather, for the highest expressed proviruses, 3q12.3 and 1q22, both show transcription originating from different repeat elements, namely an Alu element and a solo LTR from the ERV1 group of endogenous retroviruses (Fig 4-15). Transcription originating in non-HML-2 elements in PBMCs is in direct contrast to what is observed in Tera-1 cells, which exhibited robust HML-2 LTR activity driving transcription of the highest expressed HML-2 proviruses.

Relating to 3q12.3 transcription, the region on chromosome 3 where it resides is highly active, as it shows a high level of transcription of nearby gene *RPL24* in both types of PBMCs and deposition of H3K27 acetylation marks, associated with areas of active transcription [263], based on publicly available ChIP-Seq ENCODE data (Multiple cell lines, Broad Institute) [41]. The Alu sequence that appears to be driving its transcription is of the expected length for this type of element (309 bp as compared to ~280 on average) but transcription starts towards the end of the Alu element, not the

beginning where its RNA polymerase III promoters usually reside [56], indicating non-canonical transcription. Though it is notable that this provirus encodes an intact *gag* ORF, 3q12.3 was not differentially expressed between HIV-1 infected and uninfected populations (Fig 4-14). Furthermore, packaged HML-2 RNA was not detected in plasma (Fig 3-2), implying that 3q12.3 Gag protein may not be able to produce virions that package HML-2 RNA.

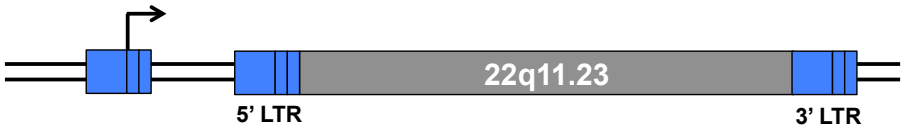
1q22 transcription is driven by a truncated LTR12F element located downstream from the provirus and whose position in the genome is associated with H3K27 acetylation as well (GM12878 lymphoblastoid cell line, Broad Institute) [41]. Splicing from this specific LTR12F element into the 1q22 HML-2 provirus was previously annotated as lincRNA TCONS_12_00002667 [30], which partially coincides with the transcript we detect as expressed in PBMCs. LincRNA TCONS_12_00002667 was noted to be highly expressed in white blood cells, as well as in breast tissue [30]. The tissue specificity of expression likely relates to transcriptional regulation of the LTR12F element. The transcription factor CEBPB, reported to bind to the LTR12F sequence in an ENCODE transcription factor data set (IMR90 lung fibroblast cells, Stanford) [41], was not upregulated in HIV-1 infection, as determined in our RNASeq study. However, this observation does not exclude its potential role in regulating expression of 1q22 from the LTR12F in our HIV-1 patient population. CEBPB is regulated by post-translational modifications and its gene targets are regulated at the proteomic level by its association with other transcription factors [174].

The consequence of increased 1q22 expression in HIV-1 infected individuals may not be significant to HIV-1 pathogenesis, as this provirus does not maintain ORFs for the

**Figure 5-1. Observed patterns of HML-2 transcription in PBMCs and Tera-1 cells.** Cartoons are shown for different modes of HML-2 transcription, as observed with the highest expressed HML-2 proviruses in PBMCs and Tera-1 cells, identified using RNASeq. For each scenario, a representative provirus for the mode of transcription is shown in parenthesis next to the description. In each cartoon, the HML-2 LTRs are shown in blue, and the retroviral genes *gag*, *pro*, *pol* and *env* are shown as a single gray block. (A) Transcription from the HML-2 proviral 5' LTR, as would be anticipated in provirus driven expression, is depicted. 22q11.21, expressed in Tera-1 cells, is noted as an example, though the proximity of the 22q11.21 provirus to the expressed gene *PRODH* is not shown. (B) Transcription from an HML-2 LTR outside of the provirus is depicted, with 22q11.23, expressed in Tera-1 cells, given as an example. (C) Transcription from a non-HML-2 repetitive element outside of the provirus is depicted, with 3q12.3 and 1q22 given as examples. 3q12.3 transcription in PBMCs originates in an upstream SINE element (AluSx3) while 1q22 transcription originates in an unrelated ERV1 LTR12F element. The proximity of the 3q12.3 provirus to the expressed gene *RPL24* is not shown. (D) Read-through transcription from a neighboring gene is depicted, with 16p13.3, primarily expressed in PBMCs from HIV-1 infected individuals, given as an example. For many of the HML-2 proviruses expressed in antisense orientation in PBMCs, the proviruses were integrated in introns, antisense to the direction of transcription (e.g. 1q23.3 in the expressed gene *CD48*, among other examples).

A Transcription from the HML-2 proviral 5’ LTR


B Transcription from an HML-2 LTR outside of the provirus




C Transcription from a non-HML-2 repetitive element outside of the provirus


D Read-through transcription from a neighboring gene

essential retroviral genes. Beyond essential retroviral genes, we did not observe spliced transcripts relating to the expression of the *np9*, which is a putative accessory gene unique to LTR Hs type 1 HML-2 proviruses like 1q22. Np9 has been postulated to have functionality relating to oncogenesis [3, 90]. However, it was recently shown that *np9* transcripts are present in both healthy and diseased tissues [29, 216]. Thus, the role of Np9 in oncogenesis remains unclear and would be difficult to prove in the absence of detailed prospective studies following individuals over time towards their progression to malignancy. In addition, the function of the Np9 protein produced specifically by 1q22 would need to be tested to verify all previous observations relating to oncogenic activity, as ORFs from different proviruses may not retain the same functionality due to accumulated mutations.

Based on our earlier studies, the lack of HML-2 RNA detection in the plasma of HIV-1 infected individuals (Fig 3-2) implies that 1q22 transcripts are not packaged into virions at a high level. This observation could be due to the lack of a functional HML-2 packaging signal on the 1q22 transcript, which may be mutated or could have been interrupted by splicing from the LTR12F element, which drives expression, into the end of the 5' LTR of 1q22 (Fig 4-15). Beyond defects in 1q22, mutations in 3q12.3 Gag, which would presumably serve as the structural component for virion assembly, could block recognition of 1q22 RNA as well.

16p13.3, which appeared to be expressed primarily in HIV-1 infected individuals, resides in an intron for the gene *FLYWCH1* and appears to be the result of read-through transcription. This fragmentary provirus does not contain a 5' LTR and there is sparse genome coverage supporting its expression. The discovery that this provirus seems to be

enriched for expression in the HIV-1 population could relate to the level of splicing activity in the infected individuals. Contrary to this hypothesis, in our RNASeq data set we did not find any significant changes in the levels of *SRSF1* or the *hnRNP* gene family, genes which participate in pre-mRNA splicing [63].

The mechanism of increased HML-2 transcription in cells remains an open issue; however, neither the levels of HIV replication nor antiretroviral medication appear to affect the level of HML-2 expression. Based on the RNASeq results, transcription of the highly expressed proviruses 3q12.3 and 1q22 appeared to be related to the active transcription occurring in other genes at their genomic locations, with 1q22 expression potentially affected by the presence of the transcription factor CEBPB. CEBPB activity is associated with response to activating stimuli like IFNγ and LPS in murine macrophages [88], and potentially could be active in the body's response to HIV-1 infection.

Immune activation in HIV-1 infection remains a possible indirect mechanism that could lead to HML-2 expression. It is known that other endogenous retroviruses, including porcine endogenous retroviruses (PERV), MLV and MMTV, exhibit increased expression after treatment with mitogens or immune activating agents [89, 122, 168, 238]. Although we did not see a difference in HML-2 upregulation between patients on or off antiretroviral therapy (Fig 3-4 and 3-5), which are patient populations with different levels of immune activation [51], the difference in immune activation was not specifically assessed in these patient groups. A previous study reported a negative correlation between immune activation (CD38+HLA-DR+) in CD4+ and CD8+ T cells from HIV-1 infected patients and HML-2 expression *in vivo* [179], though a positive effect of stimulating agents like PMA/ionomycin, PHA and IL-2 has been reported *in*

*vitro* [45, 86]. The effect of immune activation on HML-2 expression in monocytes has not been tested. These points remain to be clarified with additional studies.

*Relationship between qPCR and RNASeq studies on HML-2 expression*

The difference in total HML-2 expression between HIV-1 infected and uninfected individuals was ~1.4-fold as measured by RNASeq (Fig 4-13) and ~2-fold as measured by qPCR (Fig 3-4). The expression of provirus 1q22 in HIV-1 infected individuals appeared to be driving the overall difference in HML-2 expression between the populations based on RNASeq (Fig 4-14). There was a significant correlation between the total HML-2 expression values as measured by qPCR and RNASeq (**p=0.005, Spearman r = 0.69) for HIV-1 infected (n=9) and uninfected individuals (n=6) included in both studies. It is surprising that there is a correlation between these values because qPCR using random cDNA priming cannot discern between the orientations of the transcripts and only amplifies proviruses with the amplicon sequence, which in our case was in *env*. This method contrasts with our RNASeq study, which only considered reads uniquely aligned in the sense orientation to a provirus, and includes hits covering the entirety of the proviral sequence. There was not a significant correlation between the qPCR data and RNASeq data when considering all reads (i.e. "Unfiltered") aligning to any strand of the provirus (i.e. "Unstranded").

Both of the highest expressed proviruses 3q12.3 and 1q22 contain *env* sequences that match the primers used in qPCR analysis. In contrast, 16p13.3 does not contain the *env* sequence and thus did not contribute to the qPCR signal. When considering 1q22 and 3q12.3 FPKM individually or added together, there was no significant correlation to total

HML-2 expression as measured by qPCR. If a relationship was seen between the highest expressed HML-2 proviral FPKM and the qPCR data, it could help explain the source of the qPCR signal. The basis for the lack of correlation between 3q12.3 and 1q22 FPKM values and the qPCR data is unclear.

Based on our qPCR results on sorted cell types (Fig 3-6), monocytes appeared to trend toward higher HML-2 expression in HIV-1 infected individuals and could have been the cell type expressing increased amounts of the 1q22 provirus. There were 4 individuals in common between the sorted cell HML-2 qPCR and RNASeq studies, and while those with the highest monocyte HML-2 qPCR expression also had the highest RNASeq 1q22 FPKM out of the group ($p=0.08$, Spearman $r=1.0$), the trend is not significant and the sample size is too small to be considered conclusive. However, 1q22 expression levels should be tested using specific qPCR in sorted PBMCs to confirm, as it is possible that different cell types are each expressing 1q22 at a higher level during HIV-1 infection.

*Concluding thoughts on HML-2 expression during HIV-1 infection*

The results of this study confirmed the increased expression of HML-2 RNA in PBMCs from HIV-1 infected patients; however, they did not support the claim that HML-2 virions are present in blood. The presence of HML-2 virions, at any level, would be highly interesting. Though no HML-2 provirus has been shown to be infectious in its integrated form, studies on reconstituted HERV-K (HML-2) viruses suggest that low-level infectivity can be acquired by a few recombination events between different proviruses [60, 134]. Studies should be performed with higher sensitivity to determine if

virion production can occur and if it is correlated with HIV disease progression. However, based on our RNASeq results profiling the HML-2 proviruses, it does not appear likely that the complement of proviruses expressed would be capable of supporting virion assembly (Fig 4-16). Another, less likely, possibility to explain the difference between our findings and other reports is that HML-2 virion production occurs in a patient population not captured in our study.

There are limited clinical data profiling the exact co-morbidities and ethnicities of study subjects reported to show HML-2 virions in their plasma. Our patient population included those of varied ethnic backgrounds, and most individuals lacked co-morbidities associated with HIV, like Hepatitis C virus (HCV) or HTLV infection. The two patients in our plasma sample cohort (Table 3-1) infected with HCV did not show virion production (Fig 3-2). HTLV-1 Tax was associated with HERV-W transcription in one study [250], and HERV-K to a minor degree, however HTLV-1 infection was not associated with HML-2 specific immune responses, differing from HIV [113]. Infection with various herpesviruses has been associated with an increase in HERV transcription [127, 235, 239], although its impact on eliciting virions in vivo is unknown.

Our observations of HML-2 RNA expression in sorted PBMCs show that all the major cell types in the blood express HML-2 and possibly protein, consistent with a previous analysis using whole blood [221]. Based on these observations, it appears that use of HML-2 expression as a way to target HIV infection carries a significant risk of off-target effects. Our data suggest that HML-2 protein may be expressed in CD8+ T cells and B cells. Thus, targeting HML-2 epitopes may affect these cells and weaken the cytotoxic and humoral arms of an individual's immune response to HIV-1 infection

[261]. HML-2 expression has also been detected in embryonic stem cells, presenting an additional cause for concern [75, 92].

*Approach to HML-2 profiling in PBMCs and Tera-1 cells*

Our approach to HML-2 profiling was to create paired-end total RNA Illumina MiSeq libraries, with a maximum sequencing length of 301 bases per read. By using RNASeq instead of single genome sequencing or PCR-cloning, as performed in other studies [202], to characterize HML-2 transcription in PBMCs and the teratocarcinoma cell line Tera-1, we were able to bypass the effects of PCR primer bias in provirus amplification [202, 259], and thus achieve greater sensitivity in the breadth of proviruses identified as both expressed in cells and packaged into virions. In addition, by performing a total cellular RNA analysis, unlike other studies that focused on the deep sequencing of HML-2 *env* amplicons [26], the context of provirus expression was understood. As reported, the HML-2 transcription profiles of PBMCs and Tera-1 cells included both recent and older proviruses as well as plus strand and minus strand transcription. Furthermore, in Tera-1 cells, we discovered that the transcription of the most highly expressed LTR 5B provirus was in fact driven by an LTR Hs element upstream, while others appeared to be driven by their native LTRs or neighboring transcription units, exemplifying how RNASeq captures valuable contextual data about how HML-2 elements are expressed in specific cells.

Recently, two groups applied next-generation sequencing to address HML-2 expression in primary human lymphocytes [26, 87]. We consider our approach as combining positive attributes from both methods. One study used PacBio sequencing of HML-2 *env* amplicons generated from blood lymphocyte RNA donated from HIV-

negative patients [26]. The PacBio sequencing platform offers fewer but longer read lengths that would be superior to standard Illumina sequencing for HML-2 identification. However, the lower error rate in Illumina sequencing may offer an advantage over PacBio, and our methodology of performing total RNASeq bypasses the effects of reverse transcription and PCR primer bias in amplicon sequencing, which may select for only a subset of expressed HML-2 proviruses [26]. In comparison to the Illumina HiSeq method used in a different study to determine the HML-2 expression profile of Tat-treated blood lymphocytes from HIV-negative donors [87], our approach similarly only considers uniquely mapped reads, but takes advantage of the 3x longer read lengths available through the Illumina MiSeq platform, improving the unique identification of HML-2 proviruses. Furthermore, in our libraries, by keeping the RNA unsheared or by limiting fragmentation, we created a pool of longer library inserts, which allows for better provirus identification, especially in combination with paired-end sequencing, alignment and expression analysis as we performed, which gave up to 600 bases of coverage per sequenced fragment. Lastly, in contrast to both previous methods, our RNASeq analysis fits within the well-established TopHat-Cufflinks pipeline used in multiple fields for transcriptome analysis, offering a streamlined approach to HML-2 expression profiling.

The high sequence similarity among the recently integrated HML-2 proviruses (Fig 4-1) was predicted to complicate RNASeq analysis. Reads generated from areas of high sequence similarity can cause the phenomenon of "multi-reads," where the read will map to multiple locations in the reference genome. We did not see a large amount of multi-reads in our PBMC data sets (Fig 4-11), but, as a testament to the sequence similarity between proviruses, we found close to 50% of all reads that mapped to HML-2

proviruses were multi-reads in the Tera-1 cell RNASeq library, and ~60% in the Tera-1 virion RNASeq library (Fig 4-2). This observation speaks to the high number of recently integrated HML-2 proviruses expressed in these samples; however, it causes substantial confusion in terms of assigning the read to a specific locus. Accurate locus-specific assignment is critical for understanding the biological relevance of HML-2 expression. To circumvent this complication, we used a filter to consider only uniquely mapped reads for the transcription profile. Although the RNASeq *in-silico* simulation (Fig 4-1) showed that this approach underrepresented both human specific LTR Hs proviruses as well as known duplicated proviruses in the genome, importantly, it is still able to capture their expression, albeit at a lower level. As sequencing read lengths increase, the alignability of reads from these highly related loci will correspondingly increase and the effects of a conservative Unique Only alignment should not hamper detection of modestly expressed loci. An approach to maximize the utility of data generated from Illumina MiSeq is to custom prepare libraries so that the RNA input is not over sheared, which negatively affects the insert size available for sequencing, and also to enrich the library for longer inserts, which can be achieved using size selection during library preparation. These steps would circumvent the favored sequencing of shorter molecules during the sequencing reaction and give longer sequence for alignment in downstream analysis. In our PBMC and Tera-1 analyses, we found that the relative transcript expression values (FPKM) of the most highly transcribed proviruses did not appear to be greatly affected by this Unique Only approach, although detection of less well transcribed proviruses was reduced in the Tera-1 data set (Fig 4-2).

LTR activity assays (as performed in Fig 4-8 and 4-9) were used to ascertain whether cloned HML-2 LTRs could actively promote transcription, as well as whether the genomic context of the LTR had an affect on its activity. In addition to these two reasons, LTR activity assays were used to inform if the Unique Only approach did exclude proviruses legitimately expressed at low levels in Tera-1 cells. In this analysis, 5' LTRs were cloned and assayed for LTR activity from three recently integrated and highly similar proviruses (1p31.1a, 11q22.1 and 12q13.2, see Fig 4-1) that showed a decrease in relative expression after Unique Only analysis (Fig 4-2) and were detected at lower than 0.5% of all HML-2 reads in the hg19 alignment. LTR activity was then compared back to either the Unfiltered FPKM or the Unique Only FPKM generated for the locus. LTR activities of the selected loci appear to relate to the Unique Only expression value, supporting the idea that the unfiltered FPKM overestimated their expression, and potentially extending the Unique Only analysis even for poorly expressed loci (data not shown). However, even though the Unique FPKM and LTR activity values look similar, the exact relationship of the proviral relative transcript expression value to its LTR activity is not established and cannot be interpreted definitively.

Filtering out multi-reads can lead to gaps in read coverage for transcripts from closely related proviruses. For the LTR Hs provirus at 22q11.21, almost full proviral coverage is seen when reads are Unfiltered (Fig 4-7). However, after Unique Only selection, coverage is clearly limited to several unique portions of the genome. Another way coverage of a provirus can be interrupted is due to polymorphisms in the donor sequence that are not present in the reference. For example, a region in the *pol* gene of provirus 22q11.21 does not have substantial read coverage in either the Unfiltered or Unique Only

alignments (Fig 4-7, red arrow). Depending on the provirus, regions missing reads could indicate that there are mutations in the donor sequence that cause reads to misalign to other related proviruses, or potentially remain unaligned if too divergent. Through sequence analysis of the 22q11.21 provirus, we established the presence of 4 SNPs over ~200 bases in *pol* that overlapped the gap in read coverage. If gaps in coverage for a provirus in the unfiltered alignment are pervasive and do not correspond to the presence of SNPs in the donor sequence, the validity of read assignments to it could be called into question.

Another issue that arises in mapping HML-2 transcription is that not all known integrations are annotated in the hg19 build of the human genome. As mentioned previously, some HML-2 proviruses are insertionally polymorphic within the human population, while others are found as solo LTRs in some individuals, and full-length proviruses in others. To ensure that we captured all known proviruses, an HML-2 reference "genome" was assembled containing all known solo LTR and proviral sequences. Thus, an alignment to the HML-2 reference genome was run in parallel to the hg19 alignment to validate hits for both the PBMC and Tera-1 data sets. The expression values generated in the HML-2 reference alignments generally corroborated the proviruses found using hg19 for the PBMC data set, but differed in the detection of provirus 19q13.12a (Fig 4-10). This discrepancy aided in the discovery that the 19q13.12a hit in the hg19 alignment was not an accurate placement (Fig 4-10). In the Tera-1 cell alignment, a notable difference between the hg19 and HML-2 alignments was that a provirus not present in hg19 and insertionally polymorphic in the human population, referred to as 19p12d (empty site in hg19: 22414379-22414382), appeared to

be expressed in Tera-1 cells at a low level, ~1.7% of all expressed HML-2 sequences (data not shown). 19p12d is a type 1 provirus that by definition encodes a defective *env*, and does not contain ORFs for the other essential retroviral genes. Curiously, 19p12d has an unusually short 23bp 5' LTR, which has been observed in HML-2 proviruses previously [101]. These results corrected and enriched the RNASeq analysis and serve as proof of principle for the utility of performing a concurrent analysis.

*HML-2 expression in the teratocarcinoma cell line Tera-1*

Our RNASeq methodology was applied to the teratocarcinoma cell line Tera-1 and to the virions produced from this cell line in order to determine HML-2 expression and packaging. In contrast to PBMCs, where transcription of many HML-2 proviruses was driven by read-through transcription or unrelated repeat elements, transcription in Tera-1 cells mainly originated from 5' proviral LTRs. Based on the LTR phylogeny of expressed proviruses (Fig 4-5), the LTR Hs group of proviruses was much more highly represented than the older LTR 5A and 5B groups in the Tera-1 cell transcriptome. Furthermore, transfection assays using the various LTRs to drive expression of a luciferase reporter in Tera-1 cells showed levels of activity consistent with their reported relative transcript levels, at least for most of the LTR Hs proviruses (Fig 4-8).

An important outlier was the 5' LTR 5B of the provirus at 22q11.23, which had relatively low transcriptional activity compared to the LTR Hs located just upstream of the provirus (Fig 4-5 and 4-6). Furthermore, the activity of the upstream LTR Hs correlated with the transcription level of the 22q11.23 provirus (Fig 4-8). The lack of transcriptional activity from the 22q11.23 proviral LTR 5B could be due to its lack of GC

**Table 5-1. Observed trends in expressed HML-2 proviruses in PBMCs and Tera-1 cells.**

| | Tera-1 cells | | PBMCs from HIV-1 infected individuals | |
|---|---|---|---|---|
| | *FPKM* | *%* | *FPKM* | *%* |
| Total expressed HML-2 proviruses | 106 | 100 | 13.2 | 100 |
| LTR Hs proviruses | 80.1 | 75.6 | 11.3 | 85.8 |
| Proviruses less than 10 mya | 49.2 | 46.4 | 10.4 | 78.9 |
| Polymorphic proviruses | 10.6 | 10 | 0.1 | 0.8 |
| Proviruses with 1+ intact ORF | 74.2 | 70 | 6.2 | 47.1 |
| Proviruses expressed from an HML-2 promoter | 85 | 80.2 | 0 | 0.0 |
| Proviruses near expressed gene(s) (<50kb) | 56.3 | 53.1 | 12.7 | 96.4 |
| Proviruses in expressed genes (intronic) | 1.5 | 1.4 | 1.5 | 11.4 |

and TATA boxes, since deletion experiments showed that the region containing these elements was important for retaining transcriptional activity of the 22q11.23 LTR Hs in Tera-1 cells (Fig 4-9), although the individual contributions of each was not discerned from the assays. Potentially, the GC boxes are of greater importance as HML-2 LTRs are thought to function independently of TATA box and initiator elements [154] and a substantial loss in promoter activity was only seen when all GC boxes were removed from truncation constructs (Fig 4-9). Our data agree with previous observations that the ubiquitous transcription factors Sp1 and Sp3, which bind to GC boxes found in promoter sequences, could play a large role in regulating HML-2 promoter activity [74, 154].

HML-2 LTR promoter activity is cell type-specific and depends on a number of factors including epigenetics, transcription factor binding and proximity to other expressed genes [109, 131, 200]. For example, the highly expressed 22q11.21 provirus is situated very close to the expressed cellular gene *PRODH* (Fig 4-5 and 4-7), which may give the LTR access to transcriptional machinery and affect its transcription, although the reverse has also been proposed [234]. Similar LTR Hs elements (1p31.1a, 11q22.1 and

12q13.2) that retain promoter motifs but are located in less actively transcribed regions do not appear to be highly expressed in Tera-1 cells based on FPKM values. Interestingly, these LTRs also do not show high promoter activity in our *in vitro* LTR activity assay. This finding implies that even though these LTRs retain promoter motifs like GC and TATA boxes, they are missing additional promoter elements that are present on active LTRs that allow for their expression in Tera-1 cells. Along the same lines, the 22q11.21 5' LTR Hs showed high activity in promoter assays in Tera-1 cells (Fig 4-8), but very little activity in breast cancer cell lines (data not shown). It is likely that expression results from a disease-state or tissue-specific factor acting on the LTR. Tissue specific expression is exemplified by the LTR Hs driven transcription of the LTR 5B provirus on 22q11.23, which coincides with a lincRNA of unknown function annotated in hg19 [30] and whose expression is highest in prostate tissue, testes and ovaries, likely reflecting the tissue-specific transcriptional regulation of the ancestral HML-2 virus.

Some teratocarcinoma cell lines, called embryonic carcinoma cells (ECCs), retain pluripotency and are considered the malignant counterparts to embryonic stem cells (ESCs) [191]. Tera-1 cells are nullipotent by nature, meaning that they are unable to differentiate into any cell type, however they have been shown to retain the ability to express ESC related transcription factors like OCT4 and NANOG [188]. Recently, the activation of HERVs has been shown to occur in ESCs [145, 176, 206], including HML-2 proviruses [75, 92]. Here, we observed that many of the HML-2 proviruses reported as expressed in ESCs are expressed in Tera-1 cells, despite their lack of pluripotency. 22q11.21, one of the top expressed proviruses in Tera-1, was also shown to be the highest expressed HML-2 provirus in all ESCs and ECCs, in addition to 5q33.3 and 6q14.1 (Fig

4-3) [75]. These proviruses were also among the most frequently packaged transcripts in HML-2 virions produced by Tera-1 cells (Fig 4-3). Similarities in HML-2 expression may be reliant upon the expression of common transcription factors that are able to bind the HML-2 LTR, like OCT4 or NANOG. The importance of OCT4 to transcription from the LTR Hs in ESCs has been shown previously [92]. We observed that OCT4 and NANOG are highly expressed in Tera-1 cells and are not expressed in PBMCs, which may partially explain the differences in HML-2 expression between these cell types (OCT4 FPKM = 220.2; NANOG FPKM = 64.7). The 5' proviral LTR of 22q11.21 and the LTR Hs driving expression of the 22q11.23 provirus have a predicted OCT4 binding site in the first 200nt of LTR sequence. A putative OCT4 binding site is not present on the 22q11.23 proviral 5' LTR, potentially indicating its importance to HML-2 transcription in this cell line, amongst other requirements.

It is interesting to speculate that the lincRNA expressed by the 22q11.23 upstream LTR Hs, which encompasses the whole downstream LTR 5B proviral sequence (Fig 4-6), plays a role during embryogenesis or contributes to maintenance of the pluripotent state. Based on ENCODE RNASeq data, the 22q11.23 lincRNA is expressed in ESCs (H1-hESC cell line, Cold Spring Harbor Laboratory) [41]. Our detection of 22q11.23 provirus expression may not have been identified in previous studies of HML-2 expression in ECCs or ESCs due to primer bias, as the sequence of this LTR 5B provirus is divergent from that of the more recently integrated LTR Hs group targeted by most primer sets. While expression of this provirus/lincRNA alone would not be expected to have an effect on pluripotency, as made clear by the fact the nullipotent cell line Tera-1 expresses it, it could have a role in interacting with factors that are expressed in ESCs that are not

expressed in Tera-1. For example, in ESCs, lincRNAs driven by HERV-H LTRs were shown to have a role as scaffolds for binding of activating transcriptional complexes that regulate expression of nearby genes [145]. Also, the expression of a lincRNA necessary for pluripotency, named *lincRoR*, is expressed in Tera-1 cells to no effect, thus supporting the idea that the 22q11.23 lincRNA could have a role in pluripotency that is not functionally apparent in Tera-1 cells (*lincRoR* FPKM = 7.79) [176].

Alternatively, it is notable that the 22q11.23 provirus, which is an ancient integration estimated to be between 21-39 million years old (though this estimate may be inaccurate due to proviral recombination) [233], retains partial coding capacity for Gag, truncated by 43AA on its C-terminus. The maintenance of this ORF could indicate a role for 22q11.23 beyond a lincRNA, as there is evidence that assumed lincRNAs can be actually translated in the cell [105]. Viral particles of HML-2 origin were observed in blastocysts in a recent paper [92]; however, these virions did not appear to be mature. Potentially, incorporation of the truncated, ancient 22q11.23 Gag would prevent proper maturation of HML-2 particles produced during embryogenesis. The study of preserved ancient ORFs from 22q11.23 and other proviruses can lead to a greater understanding of their functionality in human biology.

Based on the predicted ORFs for the expressed proviruses, the majority of the HML-2 transcripts in Tera-1 cells encode *gag* (61%), including full-length and truncated forms, with *pro* (5%), *pol* ORF (5.5%) and *env* (4%) represented at much lower levels (Fig 4-3). Based on preliminary analysis, the full-length 22q11.21 Gag has functional protease cleavage sites, whereas in the truncated 22q11.23 Gag these sites are mutated [82]. Electron microscopy of Tera-1 virions shows immature particles budding from cells [16],

however the relative contributions of ineffective Gag processing, co-packaging of full-length and truncated Gag and/or lack of functional protease to this phenomenon were not determined. In terms of morphology, Tera-1 virions infrequently show Env studding [16], an observation consistent with our RNASeq data, which show only ~4% of HML-2 transcripts, originating from two expressed proviruses, to be capable of expressing Env protein. In fact, western blotting for TM shows that Env protein in Tera-1 cells is not detectable (data not shown). The Env protein produced from the 7p22.1 tandem duplicated provirus which contributed 70% of the possible *env* transcripts, has been shown to be functional, however Env encoded by the 6q14.1 locus is not [59].

Tera-1 virions have not been shown to be infectious [142]. The primary packaged genome originating from the Type 1 provirus 22q11.21 has only an ORF for *gag* [202]. Although we did observe the packaging of other HML-2 genomes that could potentially be co-packaged and lead to recombination, the defective nature of the particle structure is likely to impede a proper infection cycle, thus preventing recombination and infectious virus production. Interestingly, the genomes that are selected for packaging all originate from LTR Hs proviruses that are human specific (Fig 4-3 and 4-4). In fact, genomes derived from these proviruses are preferentially selected for packaging over other highly expressed proviruses in Tera-1 cells (Fig 4-4). Potentially, only the recently integrated proviruses retain a functional packaging signal on their genomes that allows for their enrichment into Tera-1 virions. A packaging signal for HML-2 has not been reported; however, if consistent with other retroviruses, it is likely be present in the 5' untranslated region upstream of the *gag* initiation codon [52], and perhaps extending into *gag*. A result that helps elucidate necessary elements for packaging is the absence of transcripts of

provirus 12q24.11 from virions, even though this recently integrated provirus is expressed in Tera-1 cells (Fig 4-3). While 12q24.11 has *gag* leader sequence, the total provirus only retains sequence from the start of the 5' LTR into the first ~400 nucleotides of *gag*. Potentially, sequence beyond the beginning of *gag* is necessary for the proper structure of the HML-2 packaging signal. 12q24.11 also has 5 polymorphisms in its *gag* leader in comparison to the highly packaged 22q11.21 provirus that might impair the packaging motif. The roles of these differences remain to be tested. The observation that highly expressed cellular RNAs appear to be nonspecifically packaged into HML-2 virions is consistent with other retroviruses [201].

The biological significance of HML-2 transcription in Tera-1 cells, and even more remarkably, their virion production, is not clear. Likely, HML-2 expression in these cells is purely a relic of LTR responsiveness to the transcriptional environment of the ancestral virus. Thus, the production of virions in these cells is coincidental to the proviruses with responsive LTR motifs. By analysis of HML-2 proviral transcription and selective packaging into virions, we should be able to elucidate elements of HML-2 biology that were relevant to their lifecycle as infectious retroviruses. Furthermore, in utilizing a high throughput approach independent of most PCR limitations, we can assess the full scope of HML-2 expression in the context of the cell. In the future, application of HML-2 profiling to additional healthy and diseased tissues will be of great use to help elucidate the effect and utility of HML-2 expression in the human host.

## Chapter 6: References


1. Agoni, L., A. Golden, C. Guha, and J. Lenz, *Neandertal and Denisovan Retroviruses.* Curr Biol, 2012. **22**(11): p. R437-8.
2. Armbruester, V., M. Sauter, E. Krautkraemer, E. Meese, A. Kleiman, B. Best, K. Roemer, and N. Mueller-Lantzsch, *A Novel Gene from the Human Endogenous Retrovirus K Expressed in Transformed Cells.* Clin Cancer Res, 2002. **8**(6): p. 1800-7.
3. Armbruester, V., M. Sauter, K. Roemer, B. Best, S. Hahn, A. Nty, A. Schmid, S. Philipp, A. Mueller, and N. Mueller-Lantzsch, *Np9 Protein of Human Endogenous Retrovirus K Interacts with Ligand of Numb Protein X.* J Virol, 2004. **78**(19): p. 10310-9.
4. Arts, E.J. and D.J. Hazuda, *Hiv-1 Antiretroviral Drug Therapy.* Cold Spring Harb Perspect Med, 2012. **2**(4): p. a007161.
5. Baltimore, D., *Rna-Dependent DNA Polymerase in Virions of Rna Tumour Viruses.* Nature, 1970. **226**: p. 1209-1211.
6. Bannert, N. and R. Kurth, *The Evolutionary Dynamics of Human Endogenous Retroviral Families.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 149-73.
7. Barbulescu, M., G. Turner, M.I. Seaman, A.S. Deinard, K.K. Kidd, and J. Lenz, *Many Human Endogenous Retrovirus K (Herv-K) Proviruses Are Unique to Humans.* Curr Biol, 1999. **9**(16): p. 861-8.
8. Barre-Sinoussi, F., J.C. Chermann, F. Rey, M.T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-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, 1983. **220**(4599): p. 868-71.

9. Beimforde, N., K. Hanke, I. Ammar, R. Kurth, and N. Bannert, *Molecular Cloning and Functional Characterization of the Human Endogenous Retrovirus K113.* Virology, 2008. **371**(1): p. 216-25.

10. Belshaw, R., A.L. Dawson, J. Woolven-Allen, J. Redding, A. Burt, and M. Tristem, *Genomewide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family Herv-K(Hml2): Implications for Present-Day Activity.* J Virol, 2005. **79**(19): p. 12507-14.

11. Belshaw, R., A. Katzourakis, J. Paces, A. Burt, and M. Tristem, *High Copy Number in Human Endogenous Retrovirus Families Is Associated with Copying Mechanisms in Addition to Reinfection.* Mol Biol Evol, 2005. **22**(4): p. 814-7.

12. Best, S., P. Le Tissier, G. Towers, and J.P. Stoye, *Positional Cloning of the Mouse Retrovirus Restriction Gene Fv1.* Nature, 1996. **382**: p. 826-829.

13. Bhardwaj, N. and J.M. Coffin, *Endogenous Retroviruses and Human Cancer: Is There Anything to the Rumors?* Cell Host Microbe, 2014. **15**(3): p. 255-9.

14. Bhardwaj, N., F. Maldarelli, J. Mellors, and J.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, 2014. **88**(19): p. 11108-20.

15. Bhardwaj, N., M. Montesion, F. Roy, and J.M. Coffin, *Differential Expression of Herv-K (Hml-2) Proviruses in Cells and Virions of the Teratocarcinoma Cell Line Tera-1.* Viruses, 2015. **7**(3): p. 939-968.

16. Bieda, K., A. Hoffmann, and K. Boller, *Phenotypic Heterogeneity of Human Endogenous Retrovirus Particles Produced by Teratocarcinoma Cell Lines.* J Gen Virol, 2001. **82**(Pt 3): p. 591-6.

17. Bittner, J.J., *Some Possible Effects of Nursing on the Mammary Gland Tumor Incidence in Mice.* Science, 1936. **84**(2172): p. 162.

18. Blaise, S., N. de Parseval, L. Benit, and T. Heidmann, *Genomewide Screening for Fusogenic Human Endogenous Retrovirus Envelopes Identifies Syncytin 2, a Gene Conserved on Primate Evolution.* Proc Natl Acad Sci U S A, 2003. **100**(22): p. 13013-8.

19. Blankson, J.N., J.D. Siliciano, and R.F. Siliciano, *Finding a Cure for Human Immunodeficiency Virus-1 Infection.* Infect Dis Clin North Am, 2014. **28**(4): p. 633-50.

20. Boeke, J.D. and J.S. Stoye, *Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements*, in *Retroviruses*, J.M. Coffin, S.H. Hughes, and H.E. Varmus, Editors. 1997, Cold Spring Harbor Laboratory Press: Cold Spring Harbor. p. 343-435.

21. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.* Bioinformatics, 2014. **30**(15): p. 2114-20.

22. Boller, K., O. Janssen, H. Schuldes, R.R. Tonjes, and R. Kurth, *Characterization of the Antibody Response Specific for the Human Endogenous Retrovirus Htdv/Herv-K.* J Virol, 1997. **71**(6): p. 4581-8.

23. Boller, K., K. Schonfeld, S. Lischer, N. Fischer, A. Hoffmann, R. Kurth, and R.R. Tonjes, *Human Endogenous Retrovirus Herv-K113 Is Capable of Producing Intact Viral Particles.* J Gen Virol, 2008. **89**(Pt 2): p. 567-72.

24. Boshoff, C. and R. Weiss, *Aids-Related Malignancies.* Nat Rev Cancer, 2002. **2**(5): p. 373-82.

25. Brady, T., Y.N. Lee, K. Ronen, N. Malani, C.C. Berry, P.D. Bieniasz, and F.D. Bushman, *Integration Target Site Selection by a Resurrected Human Endogenous Retrovirus.* Genes Dev, 2009. **23**(5): p. 633-42.

26. Brinzevich, D., G.R. Young, R. Sebra, J. Ayllon, S.M. Maio, G. Deikus, B.K. Chen, A. Fernandez-Sesma, V. Simon, and L.C. Mulder, *Hiv-1 Interacts with Human Endogenous Retrovirus K (Hml-2) Envelopes Derived from Human Primary Lymphocytes.* J Virol, 2014. **88**(11): p. 6213-23.

27. Bronson, D.L., E.E. Fraley, J. Fogh, and S.S. Kalter, *Induction of Retrovirus Particles in Human Testicular Tumor (Tera-1) Cell Cultures: An Electron Microscopic Study.* J Natl Cancer Inst, 1979. **63**(2): p. 337-9.

28. Buscher, K., U. Trefzer, M. Hofmann, W. Sterry, R. Kurth, and J. Denner, *Expression of Human Endogenous Retrovirus K in Melanomas and Melanoma Cell Lines.* Cancer Res, 2005. **65**(10): p. 4172-80.

29. Buscher, K., S. Hahn, M. Hofmann, U. Trefzer, M. Ozel, W. Sterry, J. Lower, R. Lower, R. Kurth, and J. Denner, *Expression of the Human Endogenous Retrovirus-K Transmembrane Envelope, Rec and Np9 Proteins in Melanomas and Melanoma Cell Lines.* Melanoma Res, 2006. **16**(3): p. 223-34.

30. Cabili, M.N., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J.L. Rinn, *Integrative Annotation of Human Large Intergenic Noncoding Rnas Reveals Global Properties and Specific Subclasses.* Genes Dev, 2011. **25**(18): p. 1915-27.

31. Cahill, S. and R. Valadez, *Growing Older with Hiv/Aids: New Public Health Challenges.* Am J Public Health, 2013. **103**(3): p. e7-e15.

32. Carbone, A., *Emerging Pathways in the Development of Aids-Related Lymphomas.* Lancet Oncol, 2003. **4**(1): p. 22-9.

33. Centers for Disease, C., *Kaposi's Sarcoma and Pneumocystis Pneumonia among Homosexual Men--New York City and California.* MMWR Morb Mortal Wkly Rep, 1981. **30**(25): p. 305-8.

34. Centers for Disease, C., *Pneumocystis Pneumonia--Los Angeles.* MMWR Morb Mortal Wkly Rep, 1981. **30**(21): p. 250-2.

35. Chattergoon, M.A., R. Latanich, J. Quinn, M.E. Winter, R.W. Buckheit, 3rd, J.N. Blankson, D. Pardoll, and A.L. Cox, *Hiv and Hcv Activate the Inflammasome in Monocytes and Macrophages Via Endosomal Toll-Like Receptors without Induction of Type 1 Interferon.* PLoS Pathog, 2014. **10**(5): p. e1004082.

36. Cillo, A.R., A. Krishnan, R.T. Mitsuyasu, D.K. McMahon, S. Li, J.J. Rossi, J.A. Zaia, and J.W. Mellors, *Plasma Viremia and Cellular Hiv-1 DNA Persist Despite Autologous Hematopoietic Stem Cell Transplantation for Hiv-Related Lymphoma.* J Acquir Immune Defic Syndr, 2013. **63**(4): p. 438-41.

37. Cillo, A.R., M.D. Sobolewski, R.J. Bosch, E. Fyne, M. Piatak, Jr., J.M. Coffin, and J.W. Mellors, *Quantification of Hiv-1 Latency Reversal in Resting Cd4+ T*

*Cells from Patients on Suppressive Antiretroviral Therapy.* Proc Natl Acad Sci U S A, 2014. **111**(19): p. 7078-83.

38. Cimarelli, A., S. Sandin, S. Hoglund, and J. Luban, *Basic Residues in Human Immunodeficiency Virus Type 1 Nucleocapsid Promote Virion Assembly Via Interaction with Rna [in Process Citation]*. J Virol, 2000. **74**(7): p. 3046-57.

39. Coffin, J.M., S.H. Hughes, and H.E. Varmus, *Intermezzo: The Interactions of Retroviruses and Their Hosts*, in *Retroviruses*, J.M. Coffin, S.H. Hughes, and H.E. Varmus, Editors. 1997, Cold Spring Harbor Laboratory Press: Cold Spring Harbor. p. 335-341.

40. Cohen, M.S., G.M. Shaw, A.J. McMichael, and B.F. Haynes, *Acute Hiv-1 Infection.* N Engl J Med, 2011. **364**(20): p. 1943-54.

41. Consortium, E.P., *An Integrated Encyclopedia of DNA Elements in the Human Genome.* Nature, 2012. **489**(7414): p. 57-74.

42. Contreras-Galindo, R., M. Gonzalez, S. Almodovar-Camacho, S. Gonzalez-Ramirez, E. Lorenzo, and Y. Yamamura, *A New Real-Time-Rt-Pcr for Quantitation of Human Endogenous Retroviruses Type K (Herv-K) Rna Load in Plasma Samples: Increased Herv-K Rna Titers in Hiv-1 Patients with Haart Non-Suppressive Regimens.* J Virol Methods, 2006. **136**(1-2): p. 51-7.

43. Contreras-Galindo, R., M.H. Kaplan, D.M. Markovitz, E. Lorenzo, and Y. Yamamura, *Detection of Herv-K(Hml-2) Viral Rna in Plasma of Hiv Type 1-Infected Individuals.* AIDS Res Hum Retroviruses, 2006. **22**(10): p. 979-84.

44. Contreras-Galindo, R., S. Almodovar-Camacho, S. Gonzalez-Ramirez, E. Lorenzo, and Y. Yamamura, *Comparative Longitudinal Studies of Herv-K and*

*Hiv-1 Rna Titers in Hiv-1-Infected Patients Receiving Successful Versus Unsuccessful Highly Active Antiretroviral Therapy.* AIDS Res Hum Retroviruses, 2007. **23**(9): p. 1083-6.

45. Contreras-Galindo, R., P. Lopez, R. Velez, and Y. Yamamura, *Hiv-1 Infection Increases the Expression of Human Endogenous Retroviruses Type K (Herv-K) in Vitro.* AIDS Res Hum Retroviruses, 2007. **23**(1): p. 116-22.

46. Contreras-Galindo, R., M.H. Kaplan, A.C. Contreras-Galindo, M.J. Gonzalez-Hernandez, I. Ferlenghi, F. Giusti, E. Lorenzo, S.D. Gitlin, M.H. Dosik, Y. Yamamura, and D.M. Markovitz, *Characterization of Human Endogenous Retroviral Elements in the Blood of Hiv-1-Infected Individuals.* J Virol, 2012. **86**(1): p. 262-76.

47. Contreras-Galindo, R., M.H. Kaplan, S. He, A.C. Contreras-Galindo, M.J. Gonzalez-Hernandez, F. Kappes, D. Dube, S.M. Chan, D. Robinson, F. Meng, M. Dai, S.D. Gitlin, A.M. Chinnaiyan, G.S. Omenn, and D.M. Markovitz, *Hiv Infection Reveals Widespread Expansion of Novel Centromeric Human Endogenous Retroviruses.* Genome Res, 2013. **23**(9): p. 1505-13.

48. Contreras-Galindo, R., M.H. Kaplan, D. Dube, M.J. Gonzalez-Hernandez, S. Chan, F. Meng, M. Dai, G.S. Omenn, S.D. Gitlin, and D.M. Markovitz, *Human Endogenous Retrovirus Type K (Herv-K) Particles Package and Transmit Herv-K-Related Sequences.* J Virol, 2015. **89**(14): p. 7187-201.

49. Cornelis, G., O. Heidmann, S.A. Degrelle, C. Vernochet, C. Lavialle, C. Letzelter, S. Bernard-Stoecklin, A. Hassanin, B. Mulot, M. Guillomot, I. Hue, T. Heidmann, and A. Dupressoir, *Captured Retroviral Envelope Syncytin Gene Associated with*

*the Unique Placental Structure of Higher Ruminants.* Proc Natl Acad Sci U S A, 2013. **110**(9): p. E828-37.

50. Crick, F., *Central Dogma of Molecular Biology.* Nature, 1970. **227**(5258): p. 561-3.

51. d'Ettorre, G., M. Paiardini, G. Ceccarelli, G. Silvestri, and V. Vullo, *Hiv-Associated Immune Activation: From Bench to Bedside.* AIDS Res Hum Retroviruses, 2011. **27**(4): p. 355-64.

52. D'Souza, V. and M.F. Summers, *How Retroviruses Select Their Genomes.* Nat Rev Microbiol, 2005. **3**(8): p. 643-55.

53. Dalgleish, A.G., P.C.L. Beverly, P.R. Clapham, D.H. Crawford, M.F. Greaves, and R.A. Weiss, *The Cd4 (T4) Antigen Is an Essential Component of the Receptor for the Aids Retrovirus.* Nature, 1984. **312**: p. 763-767.

54. de Parseval, N., V. Lazar, J.F. Casella, L. Benit, and T. Heidmann, *Survey of Human Genes of Retroviral Origin: Identification and Transcriptome of the Genes with Coding Capacity for Complete Envelope Proteins.* J Virol, 2003. **77**(19): p. 10414-22.

55. de The, G. and R. Bomford, *An Htlv-I Vaccine: Why, How, for Whom?* AIDS Res Hum Retroviruses, 1993. **9**(5): p. 381-6.

56. Deininger, P., *Alu Elements: Know the Sines.* Genome Biol, 2011. **12**(12): p. 236.

57. Denne, M., M. Sauter, V. Armbruester, J.D. Licht, K. Roemer, and N. Mueller-Lantzsch, *Physical and Functional Interactions of Human Endogenous Retrovirus Proteins Np9 and Rec with the Promyelocytic Leukemia Zinc Finger Protein.* J Virol, 2007. **81**(11): p. 5607-16.

58. Depil, S., C. Roche, P. Dussart, and L. Prin, *Expression of a Human Endogenous Retrovirus, Herv-K, in the Blood Cells of Leukemia Patients.* Leukemia, 2002. **16**(2): p. 254-9.

59. Dewannieux, M., S. Blaise, and T. Heidmann, *Identification of a Functional Envelope Protein from the Herv-K Family of Human Endogenous Retroviruses.* J Virol, 2005. **79**(24): p. 15573-7.

60. Dewannieux, M., F. Harper, A. Richaud, C. Letzelter, D. Ribet, G. Pierron, and T. Heidmann, *Identification of an Infectious Progenitor for the Multiple-Copy Herv-K Human Endogenous Retroelements.* Genome Res, 2006. **16**(12): p. 1548-56.

61. Dieffenbach, C.W. and A.S. Fauci, *Thirty Years of Hiv and Aids: Future Challenges and Opportunities.* Ann Intern Med, 2011. **154**(11): p. 766-71.

62. Doitsh, G., N.L. Galloway, X. Geng, Z. Yang, K.M. Monroe, O. Zepeda, P.W. Hunt, H. Hatano, S. Sowinski, I. Munoz-Arias, and W.C. Greene, *Cell Death by Pyroptosis Drives Cd4 T-Cell Depletion in Hiv-1 Infection.* Nature, 2014. **505**(7484): p. 509-14.

63. Dreyfuss, G., M.J. Matunis, S. Pinol-Roma, and C.G. Burd, *Hnrnp Proteins and the Biogenesis of Mrna.* Annu Rev Biochem, 1993. **62**: p. 289-321.

64. Dube, D., R. Contreras-Galindo, S. He, S.R. King, M.J. Gonzalez-Hernandez, S.D. Gitlin, M.H. Kaplan, and D.M. Markovitz, *Genomic Flexibility of Human Endogenous Retrovirus Type K.* J Virol, 2014.

65. Dupressoir, A., G. Marceau, C. Vernochet, L. Benit, C. Kanellopoulos, V. Sapin, and T. Heidmann, *Syncytin-a and Syncytin-B, Two Fusogenic Placenta-Specific*

*Murine Envelope Genes of Retroviral Origin Conserved in Muridae.* Proc Natl Acad Sci U S A, 2005. **102**(3): p. 725-30.

66. Ellerman, R. and O. Bang, *Experimentelle Leukamie Bei Huhnern.* Zebtr. Bakt. Parasit. Abt. 1, 1908. **64**: p. 595-609.

67. Emerson, R.O. and J.H. Thomas, *Adaptive Evolution in Zinc Finger Transcription Factors.* PLoS Genet, 2009. **5**(1): p. e1000325.

68. Faff, O., A.B. Murray, J. Schmidt, C. Leib-Mosch, V. Erfle, and R. Hehlmann, *Retrovirus-Like Particles from the Human T47d Cell Line Are Related to Mouse Mammary Tumour Virus and Are of Human Endogenous Origin.* J Gen Virol, 1992. **73 ( Pt 5)**: p. 1087-97.

69. Fauci, A.S. and R.C. Desrosiers, *Pathogenesis of Hiv and Siv*, in *Retroviruses*, J.M. Coffin, S.H. Hughes, and H.E. Varmus, Editors. 1997, Cold Spring Harbor Laboratory Press: Cold Spring Harbor. p. 587-635.

70. Feschotte, C. and C. Gilbert, *Endogenous Viruses: Insights into Viral Evolution and Impact on Host Biology.* Nat Rev Genet, 2012. **13**(4): p. 283-96.

71. Fevrier, M., K. Dorgham, and A. Rebollo, *Cd4 T Cell Depletion in Human Immunodeficiency Virus (Hiv) Infection: Role of Apoptosis.* Viruses, 2011. **3**(5): p. 586-612.

72. Flockerzi, A., A. Ruggieri, O. Frank, M. Sauter, E. Maldener, B. Kopper, B. Wullich, W. Seifarth, N. Muller-Lantzsch, C. Leib-Mosch, E. Meese, and J. Mayer, *Expression Patterns of Transcribed Human Endogenous Retrovirus Herv-K(Hml-2) Loci in Human Tissues and the Need for a Herv Transcriptome Project.* BMC Genomics, 2008. **9**: p. 354.

73. Freed, E.O., *Hiv-1 Replication.* Somat Cell Mol Genet, 2001. **26**(1-6): p. 13-33.

74. Fuchs, N.V., M. Kraft, C. Tondera, K.M. Hanschmann, J. Lower, and R. Lower, *Expression of the Human Endogenous Retrovirus (Herv) Group Hml-2/Herv-K Does Not Depend on Canonical Promoter Elements but Is Regulated by Transcription Factors Sp1 and Sp3.* J Virol, 2011. **85**(7): p. 3436-48.

75. Fuchs, N.V., S. Loewer, G.Q. Daley, Z. Izsvak, J. Lower, and R. Lower, *Human Endogenous Retrovirus K (Hml-2) Rna and Protein Expression Is a Marker for Human Embryonic and Induced Pluripotent Stem Cells.* Retrovirology, 2013. **10**: p. 115.

76. Galao, R.P., A. Le Tortorec, S. Pickering, T. Kueck, and S.J. Neil, *Innate Sensing of Hiv-1 Assembly by Tetherin Induces Nfkappab-Dependent Proinflammatory Responses.* Cell Host Microbe, 2012. **12**(5): p. 633-44.

77. Galli, U.M., M. Sauter, B. Lecher, S. Maurer, H. Herbst, K. Roemer, and N. Mueller-Lantzsch, *Human Endogenous Retrovirus Rec Interferes with Germ Cell Development in Mice and May Cause Carcinoma in Situ, the Predecessor Lesion of Germ Cell Tumors.* Oncogene, 2005. **24**(19): p. 3223-8.

78. Gallo, R.C., P.S. Sarin, E.P. Gelmann, M. Robert-Guroff, E. Richardson, V.S. Kalyanaraman, D. Mann, G.D. Sidhu, R.E. Stahl, S. Zolla-Pazner, J. Leibowitch, and M. Popovic, *Isolation of Human T-Cell Leukemia Virus in Acquired Immune Deficiency Syndrome (Aids).* Science, 1983. **220**(4599): p. 865-7.

79. Gamble, T.R., S. Yoo, F.F. Vajdos, U.K. von Schwedler, D.K. Worthylake, H. Wang, J.P. McCutcheon, W.I. Sundquist, and C.P. Hill, *Structure of the*

*Carboxyl-Terminal Dimerization Domain of the Hiv-1 Capsid Protein.* Science, 1997. **278**: p. 849-853.

80. Gao, F., E. Bailes, D.L. Robertson, Y. Chen, C.M. Rodenburg, S.F. Michael, L.B. Cummins, L.O. Arthur, M. Peeters, G.M. Shaw, P.M. Sharp, and B.H. Hahn, *Origin of Hiv-1 in the Chimpanzee Pan Troglodytes Troglodytes.* Nature, 1999. **397**(6718): p. 436-41.

81. Garrison, K.E., R.B. Jones, D.A. Meiklejohn, N. Anwar, L.C. Ndhlovu, J.M. Chapman, A.L. Erickson, A. Agrawal, G. Spotts, F.M. Hecht, S. Rakoff-Nahoum, J. Lenz, M.A. Ostrowski, and D.F. Nixon, *T Cell Responses to Human Endogenous Retroviruses in Hiv-1 Infection.* PLoS Pathog, 2007. **3**(11): p. e165.

82. George, M., T. Schwecke, N. Beimforde, O. Hohn, C. Chudak, A. Zimmermann, R. Kurth, D. Naumann, and N. Bannert, *Identification of the Protease Cleavage Sites in a Reconstituted Gag Polyprotein of an Herv-K(Hml-2) Element.* Retrovirology, 2011. **8**: p. 30.

83. Gessain, A., F. Barin, J.C. Vernant, and e. al., *Antibodies to Human T-Lymphotropic Virus Type I in Patients with Tropical Spastic Paraparesis.* Lancet, 1985. **2**: p. 407-410.

84. Gessain, A. and O. Cassar, *Epidemiological Aspects and World Distribution of Htlv-1 Infection.* Front Microbiol, 2012. **3**: p. 388.

85. Goedert, J.J., M.E. Sauter, L.P. Jacobson, R.L. Vessella, M.W. Hilgartner, S.F. Leitman, M.C. Fraser, and N.G. Mueller-Lantzsch, *High Prevalence of Antibodies against Herv-K10 in Patients with Testicular Cancer but Not with Aids.* Cancer Epidemiol Biomarkers Prev, 1999. **8**(4 Pt 1): p. 293-6.

86. Gonzalez-Hernandez, M.J., M.D. Swanson, R. Contreras-Galindo, S. Cookinham, S.R. King, R.J. Noel, Jr., M.H. Kaplan, and D.M. Markovitz, *Expression of Human Endogenous Retrovirus Type K (Hml-2) Is Activated by the Tat Protein of Hiv-1.* J Virol, 2012. **86**(15): p. 7790-805.

87. Gonzalez-Hernandez, M.J., J.D. Cavalcoli, M.A. Sartor, R. Contreras-Galindo, F. Meng, M. Dai, D. Dube, A.K. Saha, S.D. Gitlin, G.S. Omenn, M.H. Kaplan, and D.M. Markovitz, *Regulation of the Herv-K (Hml-2) Transcriptome by the Hiv-1 Tat Protein.* J Virol, 2014.

88. Gorgoni, B., D. Maritano, P. Marthyn, M. Righi, and V. Poli, *C/Ebp Beta Gene Inactivation Causes Both Impaired and Enhanced Gene Expression and Inverse Regulation of Il-12 P40 and P35 Mrnas in Macrophages.* J Immunol, 2002. **168**(8): p. 4055-62.

89. Greenberger, J.S., S.M. Phillips, J.R. Stephenson, and S.A. Aaronson, *Induction of Mouse Type-C Rna Virus by Lipopolysaccharide.* J Immunol, 1975. **115**(1): p. 317-20.

90. Gross, H., S. Barth, T. Pfuhl, V. Willnecker, A. Spurk, V. Gurtsevitch, M. Sauter, B. Hu, E. Noessner, N. Mueller-Lantzsch, E. Kremmer, and F.A. Grasser, *The Np9 Protein Encoded by the Human Endogenous Retrovirus Herv-K(Hml-2) Negatively Regulates Gene Activation of the Epstein-Barr Virus Nuclear Antigen 2 (Ebna2).* Int J Cancer, 2011. **129**(5): p. 1105-15.

91. Groudine, M., R. Eisenman, and H. Weintraub, *Chromatin Structure of Endogenous Retroviral Genes and Activation by an Inhibitor of DNA Methylation.* Nature, 1981. **292**(5821): p. 311-7.

92. Grow, E.J., R.A. Flynn, S.L. Chavez, N.L. Bayless, M. Wossidlo, D.J. Wesche, L. Martin, C.B. Ware, C.A. Blish, H.Y. Chang, R.A. Pera, and J. Wysocka, *Intrinsic Retroviral Reactivation in Human Preimplantation Embryos and Pluripotent Cells.* Nature, 2015. **522**(7555): p. 221-5.

93. Grulich, A.E., M.T. van Leeuwen, M.O. Falster, and C.M. Vajdic, *Incidence of Cancers in People with Hiv/Aids Compared with Immunosuppressed Transplant Recipients: A Meta-Analysis.* Lancet, 2007. **370**(9581): p. 59-67.

94. Hahn, S., S. Ugurel, K.M. Hanschmann, H. Strobel, C. Tondera, D. Schadendorf, J. Lower, and R. Lower, *Serological Response to Human Endogenous Retrovirus K in Melanoma Patients Correlates with Survival Probability.* AIDS Res Hum Retroviruses, 2008. **24**(5): p. 717-23.

95. Harris, R.S., K.N. Bishop, A.M. Sheehy, H.M. Craig, S.K. Petersen-Mahrt, I.N. Watt, M.S. Neuberger, and M.H. Malim, *DNA Deamination Mediates Innate Immunity to Retroviral Infection.* Cell, 2003. **113**(6): p. 803-9.

96. Hayward, W.S., B.G. Neel, and S.M. Astrin, *Activation of a Cellular Onc Gene by Promoter Insertion in Alv-Induced Lymphomas.* Nature, 1981. **290**: p. 475-480.

97. Heslin, D.J., P. Murcia, F. Arnaud, K. Van Doorslaer, M. Palmarini, and J. Lenz, *A Single Amino Acid Substitution in a Segment of the Ca Protein within Gag That Has Similarity to Human Immunodeficiency Virus Type 1 Blocks Infectivity of a Human Endogenous Retrovirus K Provirus in the Human Genome.* J Virol, 2009. **83**(2): p. 1105-14.

98. Hohn, O., K. Hanke, and N. Bannert, *Herv-K(Hml-2), the Best Preserved Family of Hervs: Endogenization, Expression, and Implications in Health and Disease.* Front Oncol, 2013. **3**: p. 246.

99. Hsiao, W.-L.W., S. Gattoni-Celli, and I.B. Weinstein, *Effects of 5-Azacytidine on Expression of Endogenous Retrovirus-Related Sequences in C3h 10t1/2 Cells.* J. Virol., 1986. **57**: p. 1119-1126.

100. Huang, W., L. Li, J.R. Myers, and G.T. Marth, *Art: A Next-Generation Sequencing Read Simulator.* Bioinformatics, 2012. **28**(4): p. 593-4.

101. Hughes, J.F. and J.M. Coffin, *A Novel Endogenous Retrovirus-Related Element in the Human Genome Resembles a DNA Transposon: Evidence for an Evolutionary Link?* Genomics, 2002. **80**(5): p. 453-5.

102. Hughes, J.F. and J.M. Coffin, *Human Endogenous Retrovirus K Solo-Ltr Formation and Insertional Polymorphisms: Implications for Human and Viral Evolution.* Proc Natl Acad Sci U S A, 2004. **101**(6): p. 1668-72.

103. Hughes, J.F. and J.M. Coffin, *Human Endogenous Retroviral Elements as Indicators of Ectopic Recombination Events in the Primate Genome.* Genetics, 2005. **171**(3): p. 1183-94.

104. Hunter, E., *Viral Entry and Receptors*, in *Retroviruses*, J.M. Coffin, S.H. Hughes, and H.E. Varmus, Editors. 1997: Cold Spring Harbor (NY).

105. Ingolia, N.T., G.A. Brar, N. Stern-Ginossar, M.S. Harris, G.J. Talhouarne, S.E. Jackson, M.R. Wills, and J.S. Weissman, *Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes.* Cell Rep, 2014. **8**(5): p. 1365-79.

106. Ishitsuka, K. and K. Tamura, *Human T-Cell Leukaemia Virus Type I and Adult T-Cell Leukaemia-Lymphoma.* Lancet Oncol, 2014. **15**(11): p. e517-26.

107. Jackson, R.B. and C.C. Little, *The Existence of Non-Chromosomal Influence in the Incidence of Mammary Tumors in Mice.* Science, 1933. **78**(2029): p. 465-6.

108. Jarmuz, A., A. Chester, J. Bayliss, J. Gisbourne, I. Dunham, J. Scott, and N. Navaratnam, *An Anthropoid-Specific Locus of Orphan C to U Rna-Editing Enzymes on Chromosome 22.* Genomics, 2002. **79**(3): p. 285-96.

109. Jern, P. and J.M. Coffin, *Effects of Retroviruses on Host Genome Function.* Annu Rev Genet, 2008. **42**: p. 709-32.

110. Jha, A.R., S.K. Pillai, V.A. York, E.R. Sharp, E.C. Storm, D.J. Wachter, J.N. Martin, S.G. Deeks, M.G. Rosenberg, D.F. Nixon, and K.E. Garrison, *Cross-Sectional Dating of Novel Haplotypes of Herv-K 113 and Herv-K 115 Indicate These Proviruses Originated in Africa before Homo Sapiens.* Mol Biol Evol, 2009. **26**(11): p. 2617-26.

111. Jha, A.R., D.F. Nixon, M.G. Rosenberg, J.N. Martin, S.G. Deeks, R.R. Hudson, K.E. Garrison, and S.K. Pillai, *Human Endogenous Retrovirus K106 (Herv-K106) Was Infectious after the Emergence of Anatomically Modern Humans.* PLoS One, 2011. **6**(5): p. e20234.

112. Jones, R.B., K.E. Garrison, S. Mujib, V. Mihajlovic, N. Aidarus, D.V. Hunter, E. Martin, V.M. John, W. Zhan, N.F. Faruk, G. Gyenes, N.C. Sheppard, I.M. Priumboom-Brees, D.A. Goodwin, L. Chen, M. Rieger, S. Muscat-King, P.T. Loudon, C. Stanley, S.J. Holditch, J.C. Wong, K. Clayton, E. Duan, H. Song, Y. Xu, D. SenGupta, R. Tandon, J.B. Sacha, M.A. Brockman, E. Benko, C. Kovacs,

D.F. Nixon, and M.A. Ostrowski, *Herv-K-Specific T Cells Eliminate Diverse Hiv-1/2 and Siv Primary Isolates.* J Clin Invest, 2012. **122**(12): p. 4473-89.

113. Jones, R.B., F.E. Leal, A.M. Hasenkrug, A.C. Segurado, D.F. Nixon, M.A. Ostrowski, and E.G. Kallas, *Human Endogenous Retrovirus K(Hml-2) Gag and Env Specific T-Cell Responses Are Not Detected in Htlv-I-Infected Subjects Using Standard Peptide Screening Methods.* J Negat Results Biomed, 2013. **12**: p. 3.

114. Josefsson, L., M.S. King, B. Makitalo, J. Brannstrom, W. Shao, F. Maldarelli, M.F. Kearney, W.S. Hu, J. Chen, H. Gaines, J.W. Mellors, J. Albert, J.M. Coffin, and S.E. Palmer, *Majority of Cd4+ T Cells from Peripheral Blood of Hiv-1-Infected Individuals Contain Only One Hiv DNA Molecule.* Proc Natl Acad Sci U S A, 2011. **108**(27): p. 11199-204.

115. Josefsson, L., S. von Stockenstrom, N.R. Faria, E. Sinclair, P. Bacchetti, M. Killian, L. Epling, A. Tan, T. Ho, P. Lemey, W. Shao, P.W. Hunt, M. Somsouk, W. Wylie, D.C. Douek, L. Loeb, J. Custer, R. Hoh, L. Poole, S.G. Deeks, F. Hecht, and S. Palmer, *The Hiv-1 Reservoir in Eight Patients on Long-Term Suppressive Antiretroviral Therapy Is Stable with Few Genetic Changes over Time.* Proc Natl Acad Sci U S A, 2013. **110**(51): p. E4987-96.

116. Kammerer, U., A. Germeyer, S. Stengel, M. Kapp, and J. Denner, *Human Endogenous Retrovirus K (Herv-K) Is Expressed in Villous and Extravillous Cytotrophoblast Cells of the Human Placenta.* J Reprod Immunol, 2011. **91**(1-2): p. 1-8.

117. Kapusta, A., Z. Kronenberg, V.J. Lynch, X. Zhuo, L. Ramsay, G. Bourque, M. Yandell, and C. Feschotte, *Transposable Elements Are Major Contributors to the*

*Origin, Diversification, and Regulation of Vertebrate Long Noncoding Rnas.* PLoS Genet, 2013. **9**(4): p. e1003470.

118. Karolchik, D., A.S. Hinrichs, T.S. Furey, K.M. Roskin, C.W. Sugnet, D. Haussler, and W.J. Kent, *The Ucsc Table Browser Data Retrieval Tool.* Nucleic Acids Res, 2004. **32**(Database issue): p. D493-6.

119. Kaufmann, S., M. Sauter, M. Schmitt, B. Baumert, B. Best, A. Boese, K. Roemer, and N. Mueller-Lantzsch, *Human Endogenous Retrovirus Protein Rec Interacts with the Testicular Zinc-Finger Protein and Androgen Receptor.* J Gen Virol, 2010. **91**(Pt 6): p. 1494-502.

120. Kent, W.J., C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, and D. Haussler, *The Human Genome Browser at Ucsc.* Genome Res, 2002. **12**(6): p. 996-1006.

121. Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S.L. Salzberg, *Tophat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions.* Genome Biol, 2013. **14**(4): p. R36.

122. King, L.B. and R.B. Corley, *Lipopolysaccharide and Dexamethasone Induce Mouse Mammary Tumor Proviral Gene Expression and Differentiation in B Lymphocytes through Distinct Regulatory Pathways.* Mol Cell Biol, 1990. **10**(8): p. 4211-20.

123. Koralnik, I.J., E. Boeri, W.C. Saxinger, A. Lo Monico, J. Fullen, A. Gessain, H.-G. Guo, R.C. Gallo, P. Markham, V. Kalyanaraman, V. Hirsch, J. Allan, K. Murthy, P. Alford, J.P. Slattery, S.J. O'Brien, and G. Franchini, *Phylogenetic Associations of Human and Simian T-Cell Leukemia/Lymphotropic Virus Type 1*

*Strains: Evidence for Interspecies Transmission.* J. Virol., 1994. **68**: p. 2693-2707.

124. Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B.H. Hahn, S. Wolinsky, and T. Bhattacharya, *Timing the Ancestor of the Hiv-1 Pandemic Strains.* Science, 2000. **288**(5472): p. 1789-96.

125. Kovalskaya, E., A. Buzdin, E. Gogvadze, T. Vinogradova, and E. Sverdlov, *Functional Human Endogenous Retroviral Ltr Transcription Start Sites Are Located between the R and U5 Regions.* Virology, 2006. **346**(2): p. 373-8.

126. Kurth, R. and N. Bannert, *Beneficial and Detrimental Effects of Human Endogenous Retroviruses.* Int J Cancer, 2010. **126**(2): p. 306-14.

127. Kwun, H.J., H.J. Han, W.J. Lee, H.S. Kim, and K.L. Jang, *Transactivation of the Human Endogenous Retrovirus K Long Terminal Repeat by Herpes Simplex Virus Type 1 Immediate Early Protein 0.* Virus Res, 2002. **86**(1-2): p. 93-100.

128. Lamprecht, B., K. Walter, S. Kreher, R. Kumar, M. Hummel, D. Lenze, K. Kochert, M.A. Bouhlel, J. Richter, E. Soler, R. Stadhouders, K. Johrens, K.D. Wurster, D.F. Callen, M.F. Harte, M. Giefing, R. Barlow, H. Stein, I. Anagnostopoulos, M. Janz, P.N. Cockerill, R. Siebert, B. Dorken, C. Bonifer, and S. Mathas, *Derepression of an Endogenous Long Terminal Repeat Activates the Csf1r Proto-Oncogene in Human Lymphoma.* Nat Med, 2010. **16**(5): p. 571-9, 1p following 579.

129. Langmead, B. and S.L. Salzberg, *Fast Gapped-Read Alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

130. Lauring, A.S., T.H. Lee, J.N. Martin, P.W. Hunt, S.G. Deeks, and M. Busch, *Lack of Evidence for Mtdna as a Biomarker of Innate Immune Activation in Hiv Infection.* PLoS One, 2012. **7**(11): p. e50486.

131. Lavie, L., M. Kitova, E. Maldener, E. Meese, and J. Mayer, *Cpg Methylation Directly Regulates Transcriptional Activity of the Human Endogenous Retrovirus Family Herv-K(Hml-2).* J Virol, 2005. **79**(2): p. 876-83.

132. Lawoko, A., B. Johansson, D. Rabinayaran, R. Pipkorn, and J. Blomberg, *Increased Immunoglobulin G, but Not M, Binding to Endogenous Retroviral Antigens in Hiv-1 Infected Persons.* J Med Virol, 2000. **62**(4): p. 435-44.

133. Lee, A., A. Nolan, J. Watson, and M. Tristem, *Identification of an Ancient Endogenous Retrovirus, Predating the Divergence of the Placental Mammals.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1626): p. 20120503.

134. Lee, Y.N. and P.D. Bieniasz, *Reconstitution of an Infectious Human Endogenous Retrovirus.* PLoS Pathog, 2007. **3**(1): p. e10.

135. Lee, Y.N., M.H. Malim, and P.D. Bieniasz, *Hypermutation of an Ancient Human Retrovirus by Apobec3g.* J Virol, 2008. **82**(17): p. 8762-70.

136. Lemaitre, C., F. Harper, G. Pierron, T. Heidmann, and M. Dewannieux, *The Herv-K Human Endogenous Retrovirus Envelope Protein Antagonizes Tetherin Antiviral Activity.* J Virol, 2014. **88**(23): p. 13626-37.

137. Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and S. Genome Project Data Processing, *The Sequence Alignment/Map Format and Samtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

138. Li, L., X. Deng, P. Linsuwanon, D. Bangsberg, M.B. Bwana, P. Hunt, J.N. Martin, S.G. Deeks, and E. Delwart, *Aids Alters the Commensal Plasma Virome.* J Virol, 2013. **87**(19): p. 10912-5.

139. Li, M.D., D.L. Bronson, T.D. Lemke, and A.J. Faras, *Restricted Expression of New Herv-K Members in Human Teratocarcinoma Cells.* Virology, 1995. **208**: p. 733-741.

140. Loewer, S., M.N. Cabili, M. Guttman, Y.H. Loh, K. Thomas, I.H. Park, M. Garber, M. Curran, T. Onder, S. Agarwal, P.D. Manos, S. Datta, E.S. Lander, T.M. Schlaeger, G.Q. Daley, and J.L. Rinn, *Large Intergenic Non-Coding Rna-Ror Modulates Reprogramming of Human Induced Pluripotent Stem Cells.* Nat Genet, 2010. **42**(12): p. 1113-7.

141. Lombardi, V.C., F.W. Ruscetti, J. Das Gupta, M.A. Pfost, K.S. Hagen, D.L. Peterson, S.K. Ruscetti, R.K. Bagni, C. Petrow-Sadowski, B. Gold, M. Dean, R.H. Silverman, and J.A. Mikovits, *Detection of an Infectious Retrovirus, Xmrv, in Blood Cells of Patients with Chronic Fatigue Syndrome.* Science, 2009. **326**(5952): p. 585-9.

142. Lower, R., J. Lower, H. Frank, R. Harzmann, and R. Kurth, *Human Teratocarcinomas Cultured in Vitro Produce Unique Retrovirus-Like Viruses.* J Gen Virol, 1984. **65 ( Pt 5)**: p. 887-98.

143. Löwer, R., R.R. Tönjes, C. Korbmacher, R. Kurth, and J. Löwer, *Identification of a Rev-Related Protein by Analysis of Spliced Transcripts of the Human Endogenous Retroviruses Htdv/Herv-K.* J. Virol., 1995. **69**: p. 141-149.

144. Löwer, R., J. Löwer, and R. Kurth, *The Viruses in All of Us: Characteristics and Biological Significance of Human Endogenous Retrovirus Sequences.* Proc. Natl. Acad. Sci. U.S.A., 1996. **93**: p. 5177-5184.

145. Lu, X., F. Sachs, L. Ramsay, P.E. Jacques, J. Goke, G. Bourque, and H.H. Ng, *The Retrovirus Hervh Is a Long Noncoding Rna Required for Human Embryonic Stem Cell Identity.* Nat Struct Mol Biol, 2014. **21**(4): p. 423-5.

146. Macfarlane, C.M. and R.M. Badge, *Genome-Wide Amplification of Proviral Sequences Reveals New Polymorphic Herv-K(Hml-2) Proviruses in Humans and Chimpanzees That Are Absent from Genome Assemblies.* Retrovirology, 2015. **12**: p. 35.

147. Magin, C., R. Lower, and J. Lower, *Corf and Rcre, the Rev/Rex and Rre/Rxre Homologues of the Human Endogenous Retrovirus Family Htdv/Herv-K.* J Virol, 1999. **73**(11): p. 9496-507.

148. Magin-Lachmann, C., S. Hahn, H. Strobel, U. Held, J. Lower, and R. Lower, *Rec (Formerly Corf) Function Requires Interaction with a Complex, Folded Rna Structure within Its Responsive Element Rather Than Binding to a Discrete Specific Binding Site.* J Virol, 2001. **75**(21): p. 10359-71.

149. Magiorkinis, G., R.J. Gifford, A. Katzourakis, J. De Ranter, and R. Belshaw, *Env-Less Endogenous Retroviruses Are Genomic Superspreaders.* Proc Natl Acad Sci U S A, 2012. **109**(19): p. 7385-90.

150. Maksakova, I.A., D.L. Mager, and D. Reiss, *Keeping Active Endogenous Retroviral-Like Elements in Check: The Epigenetic Perspective.* Cell Mol Life Sci, 2008. **65**(21): p. 3329-47.

151. Maldarelli, F., 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, *Hiv Latency. Specific Hiv Integration Sites Are Linked to Clonal Expansion and Persistence of Infected Cells.* Science, 2014. **345**(6193): p. 179-83.

152. Malim, M.H., J. Hauber, S.-Y. Le, J.V. Maizel, and B.R. Cullen, *The Hiv-1 Rev Trans-Activator Acts through a Structured Target Sequence to Activate Nuclear Export of Unspliced Viral Mrna.* Nature, 1989. **338**: p. 254-257.

153. Mallet, F., O. Bouton, S. Prudhomme, V. Cheynet, G. Oriol, B. Bonnaud, G. Lucotte, L. Duret, and B. Mandrand, *The Endogenous Retroviral Locus Ervwe1 Is a Bona Fide Gene Involved in Hominoid Placental Physiology.* Proc Natl Acad Sci U S A, 2004. **101**(6): p. 1731-6.

154. Manghera, M. and R.N. Douville, *Endogenous Retrovirus-K Promoter: A Landing Strip for Inflammatory Transcription Factors?* Retrovirology, 2013. **10**(1): p. 16.

155. Marchi, E., A. Kanapin, M. Byott, G. Magiorkinis, and R. Belshaw, *Neanderthal and Denisovan Retroviruses in Modern Humans.* Curr Biol, 2013. **23**(22): p. R994-5.

156. Marchi, E., A. Kanapin, G. Magiorkinis, and R. Belshaw, *Unfixed Endogenous Retroviral Insertions in the Human Population.* J Virol, 2014. **88**(17): p. 9529-37.

157. Matreyek, K.A. and A. Engelman, *Viral and Cellular Requirements for the Nuclear Entry of Retroviral Preintegration Nucleoprotein Complexes.* Viruses, 2013. **5**(10): p. 2483-511.

158. Matsui, T., D. Leung, H. Miyashita, I.A. Maksakova, H. Miyachi, H. Kimura, M. Tachibana, M.C. Lorincz, and Y. Shinkai, *Proviral Silencing in Embryonic Stem Cells Requires the Histone Methyltransferase Eset.* Nature, 2010. **464**(7290): p. 927-31.

159. Medstrand, P., L.N. van de Lagemaat, and D.L. Mager, *Retroelement Distributions in the Human Genome: Variations Associated with Age and Proximity to Genes.* Genome Res, 2002. **12**(10): p. 1483-95.

160. Mehle, A., B. Strack, P. Ancuta, C. Zhang, M. McPike, and D. Gabuzda, *Vif Overcomes the Innate Antiviral Activity of Apobec3g by Promoting Its Degradation in the Ubiquitin-Proteasome Pathway.* J Biol Chem, 2004. **279**(9): p. 7792-8.

161. Mens, H., M. Kearney, A. Wiegand, J. Spindler, F. Maldarelli, J.W. Mellors, and J.M. Coffin, *Amplifying and Quantifying Hiv-1 Rna in Hiv Infected Individuals with Viral Loads Below the Limit of Detection by Standard Clinical Assays.* J Vis Exp, 2011. **55**(55): p. e2960.

162. Mesri, E.A., E. Cesarman, and C. Boshoff, *Kaposi's Sarcoma and Its Associated Herpesvirus.* Nat Rev Cancer, 2010. **10**(10): p. 707-19.

163. Mi, S., X. Lee, X. Li, G.M. Veldman, H. Finnerty, L. Racie, E. LaVallie, X.Y. Tang, P. Edouard, S. Howes, J.C. Keith, Jr., and J.M. McCoy, *Syncytin Is a Captive Retroviral Envelope Protein Involved in Human Placental Morphogenesis.* Nature, 2000. **403**(6771): p. 785-9.

164. Michaud, H.A., M. de Mulder, D. SenGupta, S.G. Deeks, J.N. Martin, C.D. Pilcher, F.M. Hecht, J.B. Sacha, and D.F. Nixon, *Trans-Activation, Post-*

*Transcriptional Maturation, and Induction of Antibodies to Herv-K (Hml-2) Envelope Transmembrane Protein in Hiv-1 Infection.* Retrovirology, 2014. **11**: p. 10.

165. Michaud, H.A., D. SenGupta, M. de Mulder, S.G. Deeks, J.N. Martin, J.J. Kobie, J.B. Sacha, and D.F. Nixon, *Cutting Edge: An Antibody Recognizing Ancestral Endogenous Virus Glycoproteins Mediates Antibody-Dependent Cellular Cytotoxicity on Hiv-1-Infected Cells.* J Immunol, 2014. **193**(4): p. 1544-8.

166. Monde, K., R. Contreras-Galindo, M.H. Kaplan, D.M. Markovitz, and A. Ono, *Human Endogenous Retrovirus K Gag Coassembles with Hiv-1 Gag and Reduces the Release Efficiency and Infectivity of Hiv-1.* J Virol, 2012. **86**(20): p. 11194-208.

167. Monroe, K.M., Z. Yang, J.R. Johnson, X. Geng, G. Doitsh, N.J. Krogan, and W.C. Greene, *Ifi16 DNA Sensor Is Required for Death of Lymphoid Cd4 T Cells Abortively Infected with Hiv.* Science, 2014. **343**(6169): p. 428-32.

168. Moroni, C. and G. Schumann, *Mitogen Induction of Murine C-Type Viruses. Iv. Effects of Lipoprotein E. Coli, Pokeweed Mitogen and Dextran Sulphate.* J Gen Virol, 1978. **38**(3): p. 497-503.

169. Morozov, V.A., V.L. Dao Thi, and J. Denner, *The Transmembrane Protein of the Human Endogenous Retrovirus--K (Herv-K) Modulates Cytokine Release and Gene Expression.* PLoS One, 2013. **8**(8): p. e70399.

170. Mortazavi, A., B.A. Williams, K. McCue, L. Schaeffer, and B. Wold, *Mapping and Quantifying Mammalian Transcriptomes by Rna-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.

171. Muriaux, D., J. Mirro, D. Harvin, and A. Rein, *Rna Is a Structural Element in Retrovirus Particles.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5246-51.

172. Muster, T., A. Waltenberger, A. Grassauer, S. Hirschl, P. Caucig, I. Romirer, D. Fodinger, H. Seppele, O. Schanab, C. Magin-Lachmann, R. Lower, B. Jansen, H. Pehamberger, and K. Wolff, *An Endogenous Retrovirus Derived from Human Melanoma Cells.* Cancer Res, 2003. **63**(24): p. 8735-41.

173. Neil, S.J., T. Zang, and P.D. Bieniasz, *Tetherin Inhibits Retrovirus Release and Is Antagonized by Hiv-1 Vpu.* Nature, 2008. **451**(7177): p. 425-30.

174. Nerlov, C., *C/Ebps: Recipients of Extracellular Signals through Proteome Modulation.* Curr Opin Cell Biol, 2008. **20**(2): p. 180-5.

175. Odaka, T., H. Ikeda, and T. Akatsuka, *Restricted Expression of Endogenous N-Tropic Xc-Positive Leukemia Virus in Hybrids between G and Akr Mice: An Effect of the Fv-4r Gene.* Int J Cancer, 1980. **25**(6): p. 757-62.

176. Ohnuki, M., K. Tanabe, K. Sutou, I. Teramoto, Y. Sawamura, M. Narita, M. Nakamura, Y. Tokunaga, M. Nakamura, A. Watanabe, S. Yamanaka, and K. Takahashi, *Dynamic Regulation of Human Endogenous Retroviruses Mediates Factor-Induced Reprogramming and Differentiation Potential.* Proc Natl Acad Sci U S A, 2014. **111**(34): p. 12426-31.

177. Ono, M., M. Kawakami, and T. Takezawa, *A Novel Human Nonviral Retroposon Derived from an Endogenous Retrovirus.* Nucleic Acids Res, 1987. **15**(21): p. 8725-37.

178. Orendi, J.M., A.C. Bloem, J.C. Borleffs, F.J. Wijnholds, N.M. de Vos, H.S. Nottet, M.R. Visser, H. Snippe, J. Verhoef, and C.A. Boucher, *Activation and*

*Cell Cycle Antigens in Cd4+ and Cd8+ T Cells Correlate with Plasma Human Immunodeficiency Virus (Hiv-1) Rna Level in Hiv-1 Infection.* J Infect Dis, 1998. **178**(5): p. 1279-87.

179. Ormsby, C.E., D. Sengupta, R. Tandon, S.G. Deeks, J.N. Martin, R.B. Jones, M.A. Ostrowski, K.E. Garrison, J.A. Vazquez-Perez, G. Reyes-Teran, and D.F. Nixon, *Human Endogenous Retrovirus Expression Is Inversely Associated with Chronic Immune Activation in Hiv-1 Infection.* PLoS One, 2012. **7**(8): p. e41021.

180. Orzalli, M.H. and D.M. Knipe, *Cellular Sensing of Viral DNA and Viral Evasion Mechanisms.* Annu Rev Microbiol, 2014. **68**: p. 477-92.

181. Padow, M., L. Lai, R.J. Fisher, Y.C. Zhou, X. Wu, J.C. Kappes, and E.M. Towler, *Analysis of Human Immunodeficiency Virus Type 1 Containing Herv-K Protease.* AIDS Res Hum Retroviruses, 2000. **16**(18): p. 1973-80.

182. Palmarini, M., C.A. Gray, K. Carpenter, H. Fan, F.W. Bazer, and T.E. Spencer, *Expression of Endogenous Betaretroviruses in the Ovine Uterus: Effects of Neonatal Age, Estrous Cycle, Pregnancy, and Progesterone.* J Virol, 2001. **75**(23): p. 11319-27.

183. Palmarini, M., M. Mura, and T.E. Spencer, *Endogenous Betaretroviruses of Sheep: Teaching New Lessons in Retroviral Interference and Adaptation.* J Gen Virol, 2004. **85**(Pt 1): p. 1-13.

184. Palmer, S., M. Kearney, F. Maldarelli, E.K. Halvas, C.J. Bixby, H. Bazmi, D. Rock, J. Falloon, R.T. Davey, Jr., R.L. Dewar, J.A. Metcalf, S. Hammer, J.W. Mellors, and J.M. Coffin, *Multiple, Linked Human Immunodeficiency Virus Type*

*1 Drug Resistance Mutations in Treatment-Experienced Patients Are Missed by Standard Genotype Analysis.* J Clin Microbiol, 2005. **43**(1): p. 406-13.

185. Paludan, S.R. and A.G. Bowie, *Immune Sensing of DNA.* Immunity, 2013. **38**(5): p. 870-80.

186. Paprotka, T., K.A. Delviks-Frankenberry, O. Cingoz, A. Martinez, H.J. Kung, C.G. Tepper, W.S. Hu, M.J. Fivash, Jr., J.M. Coffin, and V.K. Pathak, *Recombinant Origin of the Retrovirus Xmrv.* Science, 2011. **333**(6038): p. 97-101.

187. Payne, L.N., P.K. Pani, and R.A. Weiss, *A Dominant Epistatic Gene Which Inhibits Susceptibility to Rsv (Rav-0).* J. Gen. Virol., 1971. **13**: p. 235-244.

188. Perrett, R.M., L. Turnpenny, J.J. Eckert, M. O'Shea, S.B. Sonne, I.T. Cameron, D.I. Wilson, E. Rajpert-De Meyts, and N.A. Hanley, *The Early Human Germ Cell Lineage Does Not Express Sox2 During in Vivo Development or Upon in Vitro Culture.* Biol Reprod, 2008. **78**(5): p. 852-8.

189. Plaeger, S.F., B.S. Collins, R. Musib, S.G. Deeks, S. Read, and A. Embry, *Immune Activation in the Pathogenesis of Treated Chronic Hiv Disease: A Workshop Summary.* AIDS Res Hum Retroviruses, 2012. **28**(5): p. 469-77.

190. Poiesz, B.J., F.W. Ruscetti, A.F. Gazdar, P.A. Bunn, J.D. Minna, and R.C. Gallo, *Detection and Isolation of Type C Retrovirus Particles from Fresh and Cultured Lymphocytes of a Patient with Cutaneous T-Cell Lymphoma.* Proc Natl Acad Sci U S A, 1980. **77**(12): p. 7415-9.

191. Przyborski, S.A., V.B. Christie, M.W. Hayman, R. Stewart, and G.M. Horrocks, *Human Embryonal Carcinoma Stem Cells: Models of Embryonic Development in Humans.* Stem Cells Dev, 2004. **13**(4): p. 400-8.

192. Rabson, A.B. and B.J. Graves, *Synthesis and Processing of Viral Rna*, in *Retroviruses*, J.M. Coffin, S.H. Hughes, and H.E. Varmus, Editors. 1997, Cold Spring Harbor Laboratory Press: Cold Spring Harbor. p. 205-261.

193. Reiss, D., Y. Zhang, and D.L. Mager, *Widely Variable Endogenous Retroviral Methylation Levels in Human Placenta.* Nucleic Acids Res, 2007. **35**(14): p. 4743-54.

194. Roberts, A., C. Trapnell, J. Donaghey, J.L. Rinn, and L. Pachter, *Improving Rna-Seq Expression Estimates by Correcting for Fragment Bias.* Genome Biol, 2011. **12**(3): p. R22.

195. Robinson, H.L., S.M. Astrin, A.M. Senior, and F.H. Salazar, *Host Susceptibility to Endogenous Viruses: Defective, Glycoprotein-Expressing Proviruses Interfere with Infections.* J Virol, 1981. **40**(3): p. 745-51.

196. Rosenberg, N. and P. Jolicoeur, *Retroviral Pathogenesis*, in *Retroviruses*, J.M. Coffin, S.H. Hughes, and H.E. Varmus, Editors. 1997, Cold Spring Harbor Laboratory Press: Cold Spring Harbor. p. 475-585.

197. Rous, P., *A Sarcoma of the Fowl Transmissible by an Agent Separable from the Tumor Cells.* J. Exp. Med., 1911. **13**: p. 397-411.

198. Rowe, H.M., J. Jakobsson, D. Mesnard, J. Rougemont, S. Reynard, T. Aktas, P.V. Maillard, H. Layard-Liesching, S. Verp, J. Marquis, F. Spitz, D.B. Constam, and D. Trono, *Kap1 Controls Endogenous Retroviruses in Embryonic Stem Cells.* Nature, 2010. **463**(7278): p. 237-40.

199. Rowe, H.M. and D. Trono, *Dynamic Control of Endogenous Retroviruses During Development.* Virology, 2011. **411**(2): p. 273-87.

200. Ruda, V.M., S.B. Akopov, D.O. Trubetskoy, N.L. Manuylov, A.S. Vetchinova, L.L. Zavalova, L.G. Nikolaev, and E.D. Sverdlov, *Tissue Specificity of Enhancer and Promoter Activities of a Herv-K(Hml-2) Ltr.* Virus Res, 2004. **104**(1): p. 11-6.

201. Rulli, S.J., Jr., C.S. Hibbert, J. Mirro, T. Pederson, S. Biswal, and A. Rein, *Selective and Nonselective Packaging of Cellular Rnas in Retrovirus Particles.* J Virol, 2007. **81**(12): p. 6623-31.

202. Ruprecht, K., H. Ferreira, A. Flockerzi, S. Wahl, M. Sauter, J. Mayer, and N. Mueller-Lantzsch, *Human Endogenous Retrovirus Family Herv-K(Hml-2) Rna Transcripts Are Selectively Packaged into Retroviral Particles Produced by the Human Germ Cell Tumor Line Tera-1 and Originate Mainly from a Provirus on Chromosome 22q11.21.* J Virol, 2008. **82**(20): p. 10008-16.

203. Ruprecht, K., J. Mayer, M. Sauter, K. Roemer, and N. Mueller-Lantzsch, *Endogenous Retroviruses and Cancer.* Cell Mol Life Sci, 2008. **65**(21): p. 3366-82.

204. Sacha, J.B., I.J. Kim, L. Chen, J.H. Ullah, D.A. Goodwin, H.A. Simmons, D.I. Schenkman, F. von Pelchrzim, R.J. Gifford, F.A. Nimityongskul, L.P. Newman, S. Wildeboer, P.B. Lappin, D. Hammond, P. Castrovinci, S.M. Piaskowski, J.S. Reed, K.A. Beheler, T. Tharmanathan, N. Zhang, S. Muscat-King, M. Rieger, C. Fernandes, K. Rumpel, J.P. Gardner, 2nd, D.H. Gebhard, J. Janies, A. Shoieb, B.G. Pierce, D. Trajkovic, E. Rakasz, S. Rong, M. McCluskie, C. Christy, J.R. Merson, R.B. Jones, D.F. Nixon, M.A. Ostrowski, P.T. Loudon, I.M. Pruimboom-Brees, and N.C. Sheppard, *Vaccination with Cancer- and Hiv Infection-*

*Associated Endogenous Retrotransposable Elements Is Safe and Immunogenic.* J Immunol, 2012. **189**(3): p. 1467-79.

205. Samuelson, L.C., K. Wiebauer, C.M. Snow, and M.H. Meisler, *Retroviral and Pseudogene Insertion Sites Reveal the Lineage of Human Salivary and Pancreatic Amylase Genes from a Single Gene During Primate Evolution.* Mol. Cell. Biol., 1990. **10**: p. 2513-2520.

206. Santoni, F.A., J. Guerra, and J. Luban, *Herv-H Rna Is Abundant in Human Embryonic Stem Cells and a Precise Marker for Pluripotency.* Retrovirology, 2012. **9**: p. 111.

207. Sanz-Ramos, M. and J.P. Stoye, *Capsid-Binding Retrovirus Restriction Factors: Discovery, Restriction Specificity and Implications for the Development of Novel Therapeutics.* J Gen Virol, 2013. **94**(Pt 12): p. 2587-98.

208. Sauter, M., S. Schommer, E. Kremmer, K. Remberger, G. Dölken, I. Lemm, M. Buck, B. Best, D. Neumann-Haefelin, and N. Mueller-Lantzsch, *Human Endogenous Retrovirus K10: Expression of Gag Protein and Detection of Antibodies in Patients with Seminomas.* J. Virol., 1995. **69**: p. 414-421.

209. Sawyer, S.L., M. Emerman, and H.S. Malik, *Ancient Adaptive Evolution of the Primate Antiviral DNA-Editing Enzyme Apobec3g.* PLoS Biol, 2004. **2**(9): p. E275.

210. Sayah, D.M., E. Sokolskaja, L. Berthoux, and J. Luban, *Cyclophilin a Retrotransposition into Trim5 Explains Owl Monkey Resistance to Hiv-1.* Nature, 2004. **430**(6999): p. 569-73.

211. Sayers, E.W., T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, S. Federhen, M. Feolo, I.M. Fingerman, L.Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D.J. Lipman, Z. Lu, T.L. Madden, T. Madej, D.R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T.A. Tatusova, L. Wagner, Y. Wang, W.J. Wilbur, E. Yaschenko, and J. Ye, *Database Resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2011. **39**(Database issue): p. D38-51.

212. Scadden, D.T., *Aids-Related Malignancies.* Annu Rev Med, 2003. **54**: p. 285-303.

213. Schiavetti, F., J. Thonnard, D. Colau, T. Boon, and P.G. Coulie, *A Human Endogenous Retroviral Sequence Encoding an Antigen Recognized on Melanoma by Cytolytic T Lymphocytes.* Cancer Res, 2002. **62**(19): p. 5510-6.

214. Schlesinger, S., A.H. Lee, G.Z. Wang, L. Green, and S.P. Goff, *Proviral Silencing in Embryonic Cells Is Regulated by Yin Yang 1.* Cell Rep, 2013. **4**(1): p. 50-8.

215. Schmitt, K., C. Richter, C. Backes, E. Meese, K. Ruprecht, and J. Mayer, *Comprehensive Analysis of Human Endogenous Retrovirus Group Herv-W Locus Transcription in Multiple Sclerosis Brain Lesions by High-Throughput Amplicon Sequencing.* J Virol, 2013. **87**(24): p. 13837-52.

216. Schmitt, K., K. Heyne, K. Roemer, E. Meese, and J. Mayer, *Herv-K(Hml-2) Rec and Np9 Transcripts Not Restricted to Disease but Present in Many Normal Human Tissues.* Mob DNA, 2015. **6**: p. 4.

217. Schooley, R.T., *Human Retroviruses: Aids and Other Diseases*, in *Mechanisms of Microbial Disease*, V.D. N. Cary Engleberg, and Terence S. Dermody. , Editor. 2007, Lippincott Williams and Wilkins: Baltimore, Md., and Philadelphia. p. 762.

218. Schroder, A.R., P. Shinn, H. Chen, C. Berry, J.R. Ecker, and F. Bushman, *Hiv-1 Integration in the Human Genome Favors Active Genes and Local Hotspots.* Cell, 2002. **110**(4): p. 521-9.

219. Schultz, D.C., K. Ayyanathan, D. Negorev, G.G. Maul, and F.J. Rauscher, 3rd, *Setdb1: A Novel Kap-1-Associated Histone H3, Lysine 9-Specific Methyltransferase That Contributes to Hp1-Mediated Silencing of Euchromatic Genes by Krab Zinc-Finger Proteins.* Genes Dev, 2002. **16**(8): p. 919-32.

220. Seifarth, W., B. Spiess, U. Zeilfelder, C. Speth, R. Hehlmann, and C. Leib-Mosch, *Assessment of Retroviral Activity Using a Universal Retrovirus Chip.* J Virol Methods, 2003. **112**(1-2): p. 79-91.

221. Seifarth, W., O. Frank, U. Zeilfelder, B. Spiess, A.D. Greenwood, R. Hehlmann, and C. Leib-Mosch, *Comprehensive Analysis of Human Endogenous Retrovirus Transcriptional Activity in Human Tissues with a Retrovirus-Specific Microarray.* J Virol, 2005. **79**(1): p. 341-52.

222. SenGupta, D., R. Tandon, R.G. Vieira, L.C. Ndhlovu, R. Lown-Hecht, C.E. Ormsby, L. Loh, R.B. Jones, K.E. Garrison, J.N. Martin, V.A. York, G. Spotts, G. Reyes-Teran, M.A. Ostrowski, F.M. Hecht, S.G. Deeks, and D.F. Nixon, *Strong Human Endogenous Retrovirus-Specific T Cell Responses Are Associated with Control of Hiv-1 in Chronic Infection.* J Virol, 2011. **85**(14): p. 6977-85.

223. Sharp, P.M. and B.H. Hahn, *The Evolution of Hiv-1 and the Origin of Aids.* Philos Trans R Soc Lond B Biol Sci, 2010. **365**(1552): p. 2487-94.

224. Sheppard, N.C., R.B. Jones, B.J. Burwitz, F.A. Nimityongskul, L.P. Newman, M.B. Buechler, J.S. Reed, S.M. Piaskowski, K.L. Weisgrau, P.A. Castrovinci, N.A. Wilson, M.A. Ostrowski, B. Park, D.F. Nixon, E.G. Rakasz, and J.B. Sacha, *Vaccination against Endogenous Retrotransposable Element Consensus Sequences Does Not Protect Rhesus Macaques from Sivsme660 Infection and Replication.* PLoS One, 2014. **9**(3): p. e92012.

225. Shiramizu, B., B.G. Herndier, and M.S. McGrath, *Identification of a Common Clonal Human Immunodeficiency Virus Integration Site in Human Immunodeficiency Virus-Associated Lymphomas.* Cancer Res, 1994. **54**(8): p. 2069-72.

226. Simmons, G., S.A. Glynn, A.L. Komaroff, J.A. Mikovits, L.H. Tobler, J. Hackett, Jr., N. Tang, W.M. Switzer, W. Heneine, I.K. Hewlett, J. Zhao, S.C. Lo, H.J. Alter, J.M. Linnen, K. Gao, J.M. Coffin, M.F. Kearney, F.W. Ruscetti, M.A. Pfost, J. Bethel, S. Kleinman, J.A. Holmberg, M.P. Busch, and X.S.R.W.G. Blood, *Failure to Confirm Xmrv/Mlvs in the Blood of Patients with Chronic Fatigue Syndrome: A Multi-Laboratory Study.* Science, 2011. **334**(6057): p. 814-7.

227. Simpson, G.R., C. Patience, R. Löwer, R.R. Tönjes, H.D.M. Moore, R.A. Weiss, and M.T. Boyd, *Endogenous D-Type (Herv-K) Related Sequences Are Packaged into Retroviral Particles in the Placenta and Possess Open Reading Frames for Reverse Transcriptase.* Virology, 1996. **222**: p. 451-456.

228. Steinhuber, S., M. Brack, G. Hunsmann, H. Schwelberger, M.P. Dierich, and W. Vogetseder, *Distribution of Human Endogenous Retrovirus Herv-K Genomes in Humans and Different Primates.* Hum Genet, 1995. **96**(2): p. 188-92.

229. Stevens, R.W., A.L. Baltch, R.P. Smith, B.J. McCreedy, P.B. Michelsen, L.H. Bopp, and H.B. Urnovitz, *Antibody to Human Endogenous Retrovirus Peptide in Urine of Human Immunodeficiency Virus Type 1-Positive Patients.* Clin Diagn Lab Immunol, 1999. **6**(6): p. 783-6.

230. Stoye, J.P., C. Moroni, and J. Coffin, *Virological Events Leading to Spontaneous Akr Thymomas.* J. Virol., 1991. **65**: p. 1273-1285.

231. Stoye, J.P., *Studies of Endogenous Retroviruses Reveal a Continuing Evolutionary Saga.* Nat Rev Microbiol, 2012. **10**(6): p. 395-406.

232. Stremlau, M., C.M. Owens, M.J. Perron, M. Kiessling, P. Autissier, and J. Sodroski, *The Cytoplasmic Body Component Trim5alpha Restricts Hiv-1 Infection in Old World Monkeys.* Nature, 2004. **427**(6977): p. 848-53.

233. Subramanian, R.P., J.H. Wildschutte, C. Russo, and J.M. Coffin, *Identification, Characterization, and Comparative Genomic Distribution of the Herv-K (Hml-2) Group of Human Endogenous Retroviruses.* Retrovirology, 2011. **8**(1): p. 90.

234. Suntsova, M., E.V. Gogvadze, S. Salozhin, N. Gaifullin, F. Eroshkin, S.E. Dmitriev, N. Martynova, K. Kulikov, G. Malakhova, G. Tukhbatova, A.P. Bolshakov, D. Ghilarov, A. Garazha, A. Aliper, C.R. Cantor, Y. Solokhin, S. Roumiantsev, P. Balaban, A. Zhavoronkov, and A. Buzdin, *Human-Specific Endogenous Retroviral Insert Serves as an Enhancer for the Schizophrenia-Linked Gene Prodh.* Proc Natl Acad Sci U S A, 2013. **110**(48): p. 19472-7.

235. Sutkowski, N., B. Conrad, D.A. Thorley-Lawson, and B.T. Huber, *Epstein-Barr Virus Transactivates the Human Endogenous Retrovirus Herv-K18 That Encodes a Superantigen.* Immunity, 2001. **15**(4): p. 579-89.

236. Swain, A. and J.M. Coffin, *Mechanism of Transduction by Retroviruses.* Science, 1992. **255**: p. 841-845.

237. Swanstrom, R. and J.W. Wills, *Synthesis, Assembly and Processing of Viral Proteins*, in *Retroviruses*, J.M. Coffin, S.H. Hughes, and H.E. Varmus, Editors. 1997, Cold Spring Harbor Laboratory Press: Cold Spring Harbor. p. 263-334.

238. Tacke, S.J., V. Specke, and J. Denner, *Differences in Release and Determination of Subtype of Porcine Endogenous Retroviruses Produced by Stimulated Normal Pig Blood Cells.* Intervirology, 2003. **46**(1): p. 17-24.

239. Tai, A.K., J. Luka, D. Ablashi, and B.T. Huber, *Hhv-6a Infection Induces Expression of Herv-K18-Encoded Superantigen.* J Clin Virol, 2009. **46**(1): p. 47-8.

240. Takahashi, Y., N. Harashima, S. Kajigaya, H. Yokoyama, E. Cherkasova, J.P. McCoy, K. Hanada, O. Mena, R. Kurlander, A. Tawab, R. Srinivasan, A. Lundqvist, E. Malinzak, N. Geller, M.I. Lerman, and R.W. Childs, *Regression of Human Kidney Cancer Following Allogeneic Stem Cell Transplantation Is Associated with Recognition of an Herv-E Antigen by T Cells.* J Clin Invest, 2008. **118**(3): p. 1099-109.

241. Tamura, K., G. Stecher, D. Peterson, A. Filipski, and S. Kumar, *Mega6: Molecular Evolutionary Genetics Analysis Version 6.0.* Mol Biol Evol, 2013. **30**(12): p. 2725-9.

242. Tandon, R., D. SenGupta, L.C. Ndhlovu, R.G. Vieira, R.B. Jones, V.A. York, V.A. Vieira, E.R. Sharp, A.A. Wiznia, M.A. Ostrowski, M.G. Rosenberg, and D.F. Nixon, *Identification of Human Endogenous Retrovirus-Specific T Cell Responses in Vertically Hiv-1-Infected Subjects.* J Virol, 2011. **85**(21): p. 11526-31.

243. Tchasovnikarova, I.A., R.T. Timms, N.J. Matheson, K. Wals, R. Antrobus, B. Gottgens, G. Dougan, M.A. Dawson, and P.J. Lehner, *Epigenetic Silencing by the Hush Complex Mediates Position-Effect Variegation in Human Cells.* Science, 2015.

244. Tedbury, P.R. and E.O. Freed, *The Role of Matrix in Hiv-1 Envelope Glycoprotein Incorporation.* Trends Microbiol, 2014. **22**(7): p. 372-8.

245. Telesnitsky, A. and S.P. Goff, *Reverse Transcriptase and the Generation of Retroviral DNA*, in *Retroviruses*, J.M. Coffin, S.H. Hughes, and H.E. Varmus, Editors. 1997, Cold Spring Harbor Laboratory Press: Cold Spring Harbor. p. 121-160.

246. Temin, H.M., *Homology between Rna from Rous Sarcoma Virous and DNA from Rous Sarcoma Virus-Infected Cells.* Proc Natl Acad Sci U S A, 1964. **52**: p. 323-9.

247. Temin, H.M., *Formation and Activation of the Provirus of Rna Sarcoma Virus*, in *The Biology of Large Rna Viruses*, R.D. Barry and B.W. Mahy, Editors. 1970, Academic Press: London. p. 233-244.

248. Temin, H.M. and S. Mizutani, *Rna-Dependent DNA Polymerase in Virions of Rous Sarcoma Virus.* Nature, 1970. **226**: p. 1211-1213.

249. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (Igv): High-Performance Genomics Data Visualization and Exploration.* Brief Bioinform, 2013. **14**(2): p. 178-92.

250. Toufaily, C., S. Landry, C. Leib-Mosch, E. Rassart, and B. Barbeau, *Activation of Ltrs from Different Human Endogenous Retrovirus (Herv) Families by the Htlv-1 Tax Protein and T-Cell Activators.* Viruses, 2011. **3**(11): p. 2146-59.

251. Towler, E.M., S.V. Gulnik, T.N. Bhat, D. Xie, E. Gustschina, T.R. Sumpter, N. Robertson, C. Jones, M. Sauter, N. Mueller-Lantzsch, C. Debouck, and J.W. Erickson, *Functional Characterization of the Protease of Human Endogenous Retrovirus, K10: Can It Complement Hiv-1 Protease?* Biochemistry, 1998. **37**(49): p. 17137-44.

252. Triant, V.A., *Cardiovascular Disease and Hiv Infection.* Curr HIV/AIDS Rep, 2013. **10**(3): p. 199-206.

253. Turner, G., M. Barbulescu, M. Su, M.I. Jensen-Seaman, K.K. Kidd, and J. Lenz, *Insertional Polymorphisms of Full-Length Endogenous Retroviruses in Humans.* Curr Biol, 2001. **11**(19): p. 1531-5.

254. Uchiyama, T., J. Yodoi, K. Sagawa, K. Takatsuki, and H. Uchino, *Adult T-Cell Leukemia: Clinical and Hematologic Features of 16 Cases.* Blood, 1977. **50**(3): p. 481-92.

255. Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B.C. Faircloth, M. Remm, and S.G. Rozen, *Primer3--New Capabilities and Interfaces.* Nucleic Acids Res, 2012. **40**(15): p. e115.

256. van der Kuyl, A.C., *Hiv Infection and Herv Expression: A Review.* Retrovirology, 2012. **9**: p. 6.

257. Van Dooren, S., M. Salemi, and A.M. Vandamme, *Dating the Origin of the African Human T-Cell Lymphotropic Virus Type-I (Htlv-I) Subtypes.* Mol Biol Evol, 2001. **18**(4): p. 661-71.

258. Verdonck, K., E. Gonzalez, S. Van Dooren, A.M. Vandamme, G. Vanham, and E. Gotuzzo, *Human T-Lymphotropic Virus 1: Recent Knowledge About an Ancient Infection.* Lancet Infect Dis, 2007. **7**(4): p. 266-81.

259. Vincendeau, M., I. Gottesdorfer, J.M. Schreml, A.G. Wetie, J. Mayer, A.D. Greenwood, M. Helfer, S. Kramer, W. Seifarth, K. Hadian, R. Brack-Werner, and C. Leib-Mosch, *Modulation of Human Endogenous Retrovirus (Herv) Transcription During Persistent and De Novo Hiv-1 Infection.* Retrovirology, 2015. **12**: p. 27.

260. Vogt, V.M., *Retroviral Virions and Genomes*, in *Retroviruses*, J.M. Coffin, S.H. Hughes, and H.E. Varmus, Editors. 1997, Cold Spring Harbor Laboratory Press: Cold Spring Harbor. p. 27-69.

261. Walker, B.D. and X.G. Yu, *Unravelling the Mechanisms of Durable Control of Hiv-1.* Nat Rev Immunol, 2013. **13**(7): p. 487-98.

262. Wang, T., J. Zeng, C.B. Lowe, R.G. Sellers, S.R. Salama, M. Yang, S.M. Burgess, R.K. Brachmann, and D. Haussler, *Species-Specific Endogenous Retroviruses Shape the Transcriptional Network of the Human Tumor Suppressor Protein P53.* Proc Natl Acad Sci U S A, 2007. **104**(47): p. 18613-8.

263. Wang, Z., C. Zang, J.A. Rosenfeld, D.E. Schones, A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, W. Peng, M.Q. Zhang, and K. Zhao, *Combinatorial Patterns of Histone Acetylations and Methylations in the Human Genome.* Nat Genet, 2008. **40**(7): p. 897-903.

264. Wang-Johanning, F., J. Liu, K. Rycaj, M. Huang, K. Tsai, D.G. Rosen, D.T. Chen, D.W. Lu, K.F. Barnhart, and G.L. Johanning, *Expression of Multiple Human Endogenous Retrovirus Surface Envelope Proteins in Ovarian Cancer.* Int J Cancer, 2007. **120**(1): p. 81-90.

265. Wang-Johanning, F., L. Radvanyi, K. Rycaj, J.B. Plummer, P. Yan, K.J. Sastry, C.J. Piyathilake, K.K. Hunt, and G.L. Johanning, *Human Endogenous Retrovirus K Triggers an Antigen-Specific Immune Response in Breast Cancer Patients.* Cancer Res, 2008. **68**(14): p. 5869-77.

266. Wang-Johanning, F., K. Rycaj, J.B. Plummer, M. Li, B. Yin, K. Frerich, J.G. Garza, J. Shen, K. Lin, P. Yan, S.A. Glynn, T.H. Dorsey, K.K. Hunt, S. Ambs, and G.L. Johanning, *Immunotherapeutic Potential of Anti-Human Endogenous Retrovirus-K Envelope Protein Antibodies in Targeting Breast Tumors.* J Natl Cancer Inst, 2012. **104**(3): p. 189-210.

267. Wang-Johanning, F., M. Li, F.J. Esteva, K.R. Hess, B. Yin, K. Rycaj, J.B. Plummer, J.G. Garza, S. Ambs, and G.L. Johanning, *Human Endogenous Retrovirus Type K Antibodies and Mrna as Serum Biomarkers of Early-Stage Breast Cancer.* Int J Cancer, 2014. **134**(3): p. 587-95.

268. Wen, B., H. Wu, Y. Shinkai, R.A. Irizarry, and A.P. Feinberg, *Large Histone H3 Lysine 9 Dimethylated Chromatin Blocks Distinguish Differentiated from Embryonic Stem Cells.* Nat Genet, 2009. **41**(2): p. 246-50.

269. Withers-Ward, E.S., Y. Kitamura, J.P. Barnes, and J.M. Coffin, *Distribution of Targets for Avian Retrovirus DNA Integration in Vivo.* Genes & Dev., 1994. **8**: p. 1473-1487.

270. Wolf, D. and S.P. Goff, *Trim28 Mediates Primer Binding Site-Targeted Silencing of Murine Leukemia Virus in Embryonic Cells.* Cell, 2007. **131**(1): p. 46-57.

271. Wolf, D., F. Cammas, R. Losson, and S.P. Goff, *Primer Binding Site-Dependent Restriction of Murine Leukemia Virus Requires Hp1 Binding by Trim28.* J Virol, 2008. **82**(9): p. 4675-9.

272. Wolf, G., P. Yang, A.C. Fuchtbauer, E.M. Fuchtbauer, A.M. Silva, C. Park, W. Wu, A.L. Nielsen, F.S. Pedersen, and T.S. Macfarlan, *The Krab Zinc Finger Protein Zfp809 Is Required to Initiate Epigenetic Silencing of Endogenous Retroviruses.* Genes Dev, 2015. **29**(5): p. 538-54.

273. Wu, X., Y. Li, B. Crise, and S.M. Burgess, *Transcription Start Regions in the Human Genome Are Favored Targets for Mlv Integration.* Science, 2003. **300**(5626): p. 1749-51.

274. Yang, J., H.P. Bogerd, S. Peng, H. Wiegand, R. Truant, and B.R. Cullen, *An Ancient Family of Human Endogenous Retroviruses Encodes a Functional Homolog of the Hiv-1 Rev Protein.* Proc Natl Acad Sci U S A, 1999. **96**(23): p. 13404-8.

275. Yao, S., T. Sukonnik, T. Kean, R.R. Bharadwaj, P. Pasceri, and J. Ellis, *Retrovirus Silencing, Variegation, Extinction, and Memory Are Controlled by a Dynamic Interplay of Multiple Epigenetic Modifications.* Mol Ther, 2004. **10**(1): p. 27-36.

276. Yarchoan, R., G. Tosato, and R.F. Little, *Therapy Insight: Aids-Related Malignancies--the Influence of Antiviral Therapy on Pathogenesis and Management.* Nat Clin Pract Oncol, 2005. **2**(8): p. 406-15; quiz 423.

277. Yoder, J.A., C.P. Walsh, and T.H. Bestor, *Cytosine Methylation and the Ecology of Intragenomic Parasites.* Trends Genet, 1997. **13**(8): p. 335-40.

278. Young, G.R., U. Eksmond, R. Salcedo, L. Alexopoulou, J.P. Stoye, and G. Kassiotis, *Resurrection of Endogenous Retroviruses in Antibody-Deficient Mice.* Nature, 2012. **491**(7426): p. 774-8.

279. Young, G.R., J.P. Stoye, and G. Kassiotis, *Are Human Endogenous Retroviruses Pathogenic? An Approach to Testing the Hypothesis.* Bioessays, 2013. **35**(9): p. 794-803.

280. Zeilfelder, U., O. Frank, S. Sparacio, U. Schon, V. Bosch, W. Seifarth, and C. Leib-Mosch, *The Potential of Retroviral Vectors to Cotransfer Human Endogenous Retroviruses (Hervs) from Human Packaging Cell Lines.* Gene, 2007. **390**(1-2): p. 175-9.

281. Zhao, T. and M. Matsuoka, *Hbz and Its Roles in Htlv-1 Oncogenesis.* Front Microbiol, 2012. **3**: p. 247.