A Study of Polymorphic Endogenous Retroviruses in Humans: Implications in Host Health and Genome Evolution

A dissertation

submitted by

Julia Halo Wildschutte

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in Molecular Microbiology

TUFTS UNIVERSITY

Sackler School of Graduate Biomedical Sciences

August 11th of 2011

ADVISOR: John M. Coffin, Ph.D.

Copyright © by Julia Halo Wildschutte

2011

ABSTRACT

Human endogenous retroviruses (HERVs) result from the integration of retroviral DNA into germline cells. The recently active human MMTV-like HML-2 group has been the focus of this project. An analysis of the HML-2 group reveals their representation in the published genome by ~90 full-length proviruses and ~1000 solitary-LTRs. The HML-2 proviruses can be divided into three subgroups, based on the phylogenetic comparison of the paired LTRs belonging to each element, namely, the LTR5Hs, LTR5A, and LTR5B. The analyses provide sequence-based and phylogenetic support that the LTR5B are the oldest of the HML-2 proviruses, with a few shared sites in Old World monkeys. Also supported by phylogenetic analysis and by estimated times of the activities of each subgroup, the LTR5B subgroup is ancestral to both the LTR5A and LTR5-Hs. Only the LTR5-Hs have been active in humans. Phylogenetic analyses indicate the evolutionary activities of the LTR5-Hs have differed from that of the LTR5A and B subgroups, the latter two having experienced increases in copy number as a direct result of association with segmental duplication events within the host genome. Also we present the findings that the LTR5A elements are associated with a specific group of genome segments, and are found in more than half of the recently formed duplicons.

The HML-2 group represents the only HERVs with human-specific integrations, of which at least 11 HML-2 proviruses are polymorphic in integration site and frequency within the population. The expression of HML-2 proviruses has been well-reported to be up-regulated in a number of diseases. The effects of HERV expression in human tissues are poorly understood, as is how HML-2 expression relates to a particular diseased state. In this respect, polymorphic integrations have been given attention as they are conserved in sequence though present in a fraction of the population. We used high-

resolution DNA hybridization from human genomic DNAs in order to visually infer the distribution of newly formed and as-yet-uncharacterized polymorphic HML-2 proviruses. The case-control analyses of DNAs from two human diseases revealed no statistical support for a genetic association of any 'new' polymorphic element with disease, and are consistent with previous studies of similar approach.

Nevertheless, we have provided evidence that as many as 15 polymorphic proviruses are detected within human DNAs that vary in frequencies among the samples. We observed a few present at quite low frequencies, for example in one or two of 120 tested samples –below 1%, with regard to the samples analyzed, a far lower representation than seen in any other polymorphic provirus. Such a provirus would be highly conserved and might also exhibit retained functions.

ACKNOWLEDGEMENTS

I would like to express my most heartfelt thanks and gratitude to those who have offered so much help and support during my graduate work.

Most importantly, I thank my advisor, John Coffin, for his enthusiastic support of this work, and for allowing me to extend the project over different avenues to explore the observations that I found to be most interesting. John has fostered my graduate training with a style that has offered much guidance and advice, with an independence that, to me, demonstrated his respect for this work. I am grateful for his time and effort for the success of my graduate work, and look forward to a career that will benefit from this experience.

I give particular thanks to my Thesis Committee, Carol Kumamoto, Naomi Rosenberg, and Katya Heldwein, who have each given valuable time and effort to help guide my graduate work and development as a scientist. I am truly grateful to have the expertise and advice of each member of my committee, and have been fortunate for the balance and support they have offered as a collective group. I also give much thanks to my outside examiner, Paul Bieniasz, who has generously provided his time to read my thesis and travel to Tufts to lend his expertise in the field as a member of my committee.

I thank the members of the Coffin Lab, both present and past, who have always been a source of advice and knowledge, especially in coming to an unfamiliar project and topic. I would like to thank those members who have contributed specifically to this project: Walter Lech, Patric Jern, and in particular Jennifer Hughes, whose thesis has been a basis for much of the project. I thank Ravi Subramanian, who has been a colleague and friend, and one of few members of 'Team HERV-K' in the lab. Ravi's presence in the lab has led to many fruitful discussions and to collaboration, and I am grateful for the conversation and time.

I have been fortunate to have had a group of talented and inspirational students, who devoted many hours and help in this work: Crystal Russo, Daniel Ram, and Tyler Bradstreet.

I would like to thank the administrative staff, who are always so generous to offer help without hesitation. In particular I thank Verna Manni, who has devoted much time, resource, and skills for various applications, and who has become a caring and understanding friend.

The American Cancer Society and the Stanley Medical Research Institute generously provided human genomic DNA samples. Some of this work was also made possible from a National Research Service Award fellowship from the National Institutes of Mental Health.

There are friends that deserve much thanks. I thank Faith, who has been a great classmate and a friend. I will miss our science [and not] conversations, lunches, late lunches, and drinks. I thank Derick, whom I will miss for many of the same reasons, and who has been a wonderful friend to have here at Tufts. Aletheia, my best friend of more than twenty

years, through it all. My friends at home –who have truly become great friends, and I appreciate them more so than they may ever know– Niki and Jonathan. I thank Howard Stern for the mornings, and Lonnie Newburn for the late afternoons.

Finally, I owe a great amount of gratitude to my family. To my mother, who is eternally supporting of my choices, my work, and my life. Through several hard years of family illnesses and losses, she has always offered the time to talk, the needed support, and many prayers. To my father, who is equally supportive, and who has and generously provided every opportunity and means for my success. I have been extremely fortunate to have their care, guidance, and love. I thank my sister, Olivia, who is one of my best friends, confidents, and partner in crime. I am extremely appreciative of her knack at offering the best and most thoughtful advice, pointed in such a way and the 'silver lining' is actually visible. I thank my Grandmother, who is watching from elsewhere, but who gave me much loving advice, and who was one of my closest friends. Finally, I thank Hans, who offers unconditional and selfless love, attention, and support. He is my champion. I am a better person because he is a part of my life, and I would not have realized my possibilities or accomplishments without him

TABLE OF CONTENTS

CHAPTER 2

CHAPTER 3

Results and Analysis I:

Identification, characterization, and comparative genomic analysis of the HERV-K (HML-2) group of human endogenous retroviruses.

CHAPTER 4

Results and Analysis II:

Distribution of polymorphic HERV-K (HML-2) in human health and disease.

CHAPTER 5

LIST OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

LIST OF FIGURES

CHAPTER 1

CHAPTER 3

Results and Analysis I:

CHAPTER 4

Results and Analysis II:

CHAPTER 1

Introduction

Transposable elements in the human genome

Transposable elements (TEs), or "jumping genes" are discrete segments of DNA that have the ability to move, or transpose, to additional sites within the host genome [1]. Consequently, such elements have the power to affect their surrounding genomic environment, and to alter the genome structure and function of the host. About half of the human genome has been derived from repetitive elements, the vast majority of which (42%) has been contributed by the TEs [2]. Given the difficulties in the identification of ancient TEs due to their degeneration over time, and because many repeat-rich regions are not annotated within even the most recently assembled genome builds, their genomic representation is considered to be an underestimate [3].

DNA transposons

TEs can be divided into two classes, as defined by the mechanism of transposition: those that mobilize through an RNA intermediate (Class I), and those that do not (Class II) (Figure 1.1) [3]. The DNA transposons, or Class II TEs, represent less than 3% of the sequenced genome [2]. Class II TEs have been characterized in both prokaryotes and eukaryotes, and have been observed in nearly every tested species. Their widespread distribution is due in part to horizontal transmission (discussed below), and having persisted vertically through the speciation events of multiple hosts [4].

DNA transposons are considered simple transposable elements and move from site to site by a 'cut and paste' type mechanism, during which they self-excise and re-insert at distinct genomic sites while remaining in the DNA state [5]. Structurally, all

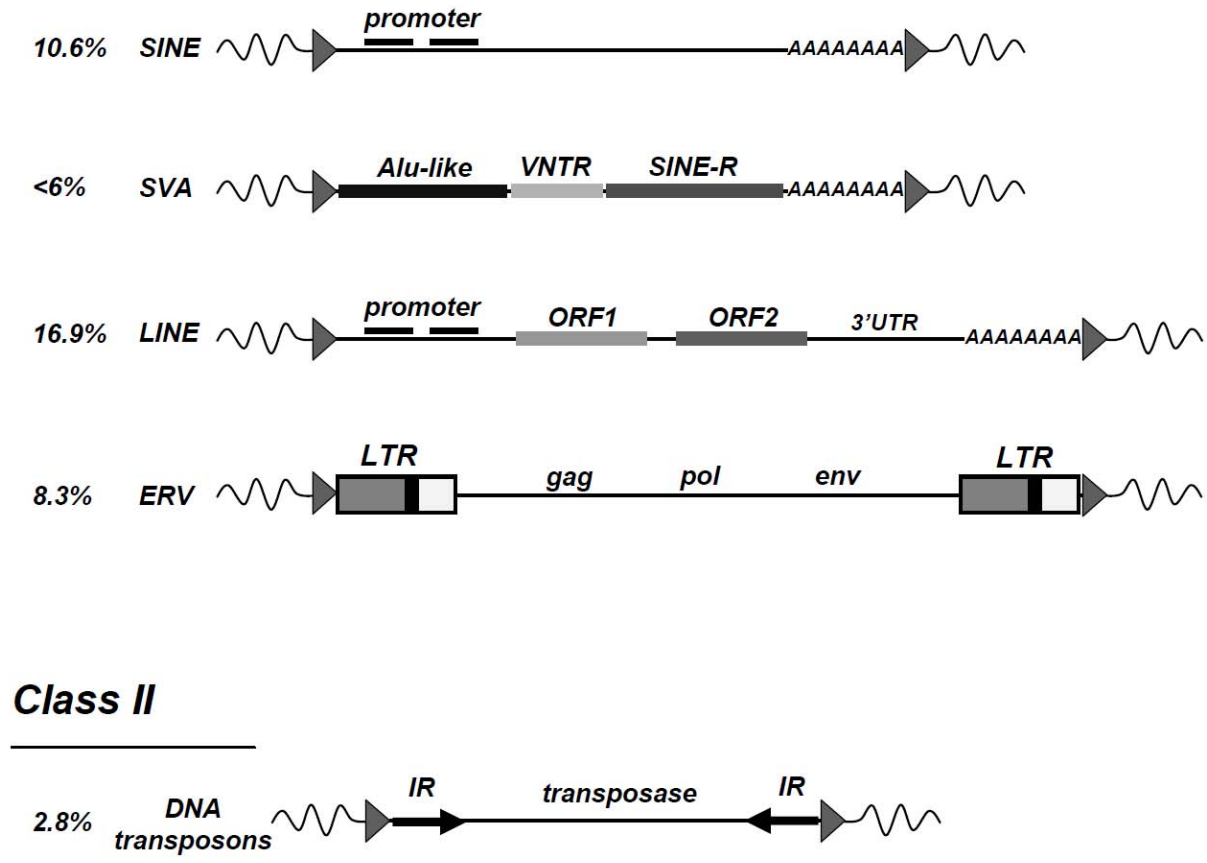


Figure 1.1. Representatives from the Class I and Class II transposable elements in human DNA.
The class I elements are divided into the non-LTR (SINEs, SVAs, and LINEs) and LTR (ERVs) retroelements. The class II elements are represented by the DNA transposons. Generalized representatives of each class are shown within host DNA and labeled according to characteristic features. At left, the relative contribution of each type within the human genome is provided as a percentage [2]. Gray triangles indicate target site duplications. The opposing arrows in the DNA transposon indicate flanking inverted repeats (IR). Abbreviations are: ORF, open reading frame; LTR, long terminal repeat; VNTR, variable number tandem repeats; UTR, untranslated region. LTRs are shown with U3 in gray, R in black, and U5 in white.

Class II TEs share a few common features. For example, they tend to be small (~1 to 2.5 kb in length) and encode only the functions necessary for their movement. More specifically, each element encodes a gene for the transposase enzyme, and a second gene that regulates its transcription. Transposase is responsible for the excision of the DNA element from the host genome, and the integration of the element into a new site. The TE is recognized by transposase by virtue of a set of inverted repeats located at either end of the element, and these differ in length and sequence depending on the DNA transposon.

The Class II DNA transposons have not been active during recent human evolution, but were evidently active in primates ~37 million years ago (mya), prior to the common ancestor of humans and New World monkeys [6]. Of 9 'superfamilies' of Class II TEs described in eukaryotes, 7 are represented in the human genome in >380,000 total copies in more than 125 'families', each with a copy number of less than 100 [7]. At least one study has characterized the evolutionary activities of these 125 groups in placental mammals (eutherians) by comparing the species distribution of DNA transposons by virtue of their insertion within other repetitive elements such as *Alu* and L1. The analyses of such *Alu* and L1 elements in multiple species led to the identification of about 40 groups that appeared to be primate specific, though without any detectable activity following the emergence of the Old World monkeys.

Although the DNA transposons are predominantly spread by vertical transmission, the phylogenies of some groups of such elements support their additional movement between host species via horizontal transfer [8, 9], for example via a parasitic vector [10]. This type of transfer is thought to provide a means of escape from extinction by accumulated mutation, as non-functional copies may be propagated *in trans* by

functional transposases, and dominant negative transposases, or the overexpression of non-functional transposase enzymes, can limit the activity of the DNA transposon. Thus, in the absence of movement to a new host species, the elements face eventual inactivation and stochastic loss of function within the host species genome [8, 9]. In this context, a suggested explanation for the apparent extinction of DNA transposons within primates ~37 mya could be either from the sudden inability for DNA transposons to be transferred to a new host, with the concomitant loss of function from mutation, or alternatively from the prevention of the transfer by the host, analogous to mechanisms of retroviral restriction [6, 11].

Non-LTR retroelements

The Class I TEs, retroelements, are transposable elements whose replication involves an RNA intermediate [12]. Retroelements differ from the DNA transposons in their properties of species distribution and transposition. Retroelements appear restricted to eukaryotes, and are moved to other genomic sites as copies, with the outcome being the replication of the elements and an increased copy number within the host, in contrast to the 'cut and paste' mechanism utilized by DNA transposons. The transposition is two-step, in which transcription of these elements generates an RNA molecule (DNA→RNA), which is converted back to a double-stranded DNA copy (RNA→DNA). The latter reaction generates a double-stranded DNA copy of the element and is catalyzed by RNA-directed DNA polymerase, or reverse transcriptase (RT) [13, 14]. The resulting DNA is inserted into another site within the host genome, completing the replication of the element. Replicative movement has potential detrimental consequences to the host;

however, most retroelements are effectively neutral to the host with little, if any, functional relevance. This observation has led to the hypothesis that these elements are inherently parasitic, or 'selfish', and will continue to amplify so long as they remain tolerable to the host [15, 16].

Collectively, retroelements have contributed to at least 42% of the human genome [2]. These are generally divided into two major classes, by virtue of the presence of long terminal repeats, or LTRs. Non-LTR retroelements are comprised of three major classes, represented by the long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and SVAs (so-named for their composite parts: SINE-R (HERV-K10-like), VNTR (variable number of tandem repeats), and an *Alu*-like region. In humans, the LTR retroelements are represented by elements derived from endogenous retroviruses (ERVs) (Figure 1.1) [12].

The SINEs, which are typified by the *Alu* family of elements, are the most abundant retroelements within the genome, totaling in excess of 1.5 million copies and corresponding to roughly 13% of the genome [2], particularly notable given their emergence within ~65 mya ago and apparent primate-specificity [17]. Structurally, *Alu* elements range from ~100-300 base pairs (bp) in length, and do not encode an open reading frame (ORF). A full-length *Alu* contains a promoter site for the cellular RNA polymerase. However, SINEs are non-autonomous, and their transcripts must be reverse transcribed and re-inserted by reverse transcriptase machinery used by other retroelements, for example from the LINEs [17]. SVAs are also non-autonomous, and so must also be complemented *in trans*. Structurally, they are similar to the *Alu* in that they are short and do not encode an ORF. However, SVAs also have no apparent promoter

sequence, and may be transcribed from 'nearby' promoter activities. Owing to their origin within the last ~25 myr, they are the least abundant retroelement at ~3000 copies and ~3% of the genome [2]. Though non-autonomous, both *Alu* and SVAs are currently mobilized within humans, at least in part by 'piggybacking' along with L1 RT [17].

The LINEs are second in abundance to SINEs, represented by about 870,000 copies, but have contributed to a greater bulk of the genome (~20%) given their average size of ~6-9 kilobases (kb). These elements have evidence of activity for ~160 mya [2]. In contrast to SINEs, LINEs are autonomous and at least some encode the proteins required for their replication. Structurally, they house a promoter for cellular RNA polymerase, ORFs (including reverse transcriptase), and a 3' untranslated region (UTR). While the vast majority of LINEs have been truncated and are replication-defective, a fraction (roughly 40-50%) still encode apparently functional RT, some of which have retained the ability for *in vitro* replication [18]. It has been estimated that 1 in every 10-100 individuals will have an additional element as a result of this activity in the germline [19]. In turn, LINE activity has implications for the germline amplification of SINEs and SVA elements, support for which has been the existence of polymorphic members of all three types (and particularly *Alu* and LINE-1) within humans [20, 21]. Due to a staggered break in the target DNA for insertion, all non-LTR retroelements are flanked by characteristic target site duplications (TSDs) anywhere from 6-20 bp [17].

LTR retroelements

The remaining 8% of retroelements in the human genome are from the LTR-containing class [2]. Represented within this class are the endogenous retroviruses

(ERVs) that have been predominantly amplified through germline infections of exogenous retroviruses [12]. The LTR elements are divided among sequences belonging to the non-autonomous MalR (~3%), full-length or near full-length ERV elements (~5%), and to ERV-derived solitary LTRs (referred to as solo LTRs), which are the products of recombination events between LTRs (~2%) [2]. Within humans, ERVs (HERVs) are categorized into one of three classes (I, II, or III) based on phylogenetic analysis of a highly conserved motif within RT [22]. Classes I and III represent older HERVs and are fixed among humans, present in ~110,000 and 83,000 copies, respectively. The Class II HERVs have formed most recently, and are present in ~8,000 copies [2].

As with the non-LTR retroelements described above, a small number of Class II HERVs are polymorphic within humans (~15 described). These elements, belonging specifically to the HERV-K(HML-2) group, are the subject of this thesis. As a whole, the HERV-K are the predominant members of the Class II ERVs, and represent the most recently formed retroviral insertions in the human germline [22]. The HML-2 have shared integration sites in the genomes of apes and Old World monkeys, but are not found in New World monkeys. Given the accepted dates for the divergence from the human to the common ancestor from either group, we can infer the appearance of the HML-2 at ~25 mya to ~35 mya [23]. Some recently formed members of HML-2 have retained ORFs, human-specific members (~25 known), and may were actively replicating within the last ~200,000 years. Collectively, these observations suggest the possibility that this group may still have activity within the human population.

Retroviruses

Overview

A retrovirus is an enveloped RNA virus whose virions carry two copies of a single-stranded positive sense (+) RNA genome [24-27]. Each genome copy ranges from ~7 to 12 kb, depending on the retrovirus type, and structurally resembles a eukaryotic messenger RNA (mRNA) transcript. For all retroviruses, the genome encodes four essential genes whose expression generates the structural and enzymatic proteins necessary for infection and replication: *gag*, *pro*, *pol*, and *env*. Retroviruses encoding only these genes are known as 'simple' retroviruses, distinguished from those encoding additional accessory genes, termed 'complex'.

A prototypical retroviral particle is diagrammed in Figure 1.2. Briefly, the structural core of the retroviral particle is made up of the nucleocapsid (NC), capsid (CA), and matrix (MA) proteins, which are originally expressed as a polyprotein of Gag but later processed during particle maturation. Dimerized RNA genome copies are tightly bound by NC in a ribonucleoprotein complex, which is in turn embedded within a capsid 'shell' made up of CA proteins. Included within the capsid are the virally-encoded enzymes RT and integrase (IN), in addition to specific host transfer RNA (tRNA) molecules which function to prime minus strand synthesis during reverse transcription of the RNA genome. A bi-lipid membrane layer acquired during budding from the host cell envelopes the capsid and associates with the particle by interaction of MA protein. The membrane facilitates the display of the viral transmembrane (TM) and surface unit (SU) Env glycoproteins that mediate receptor binding and entry into a new host cell.

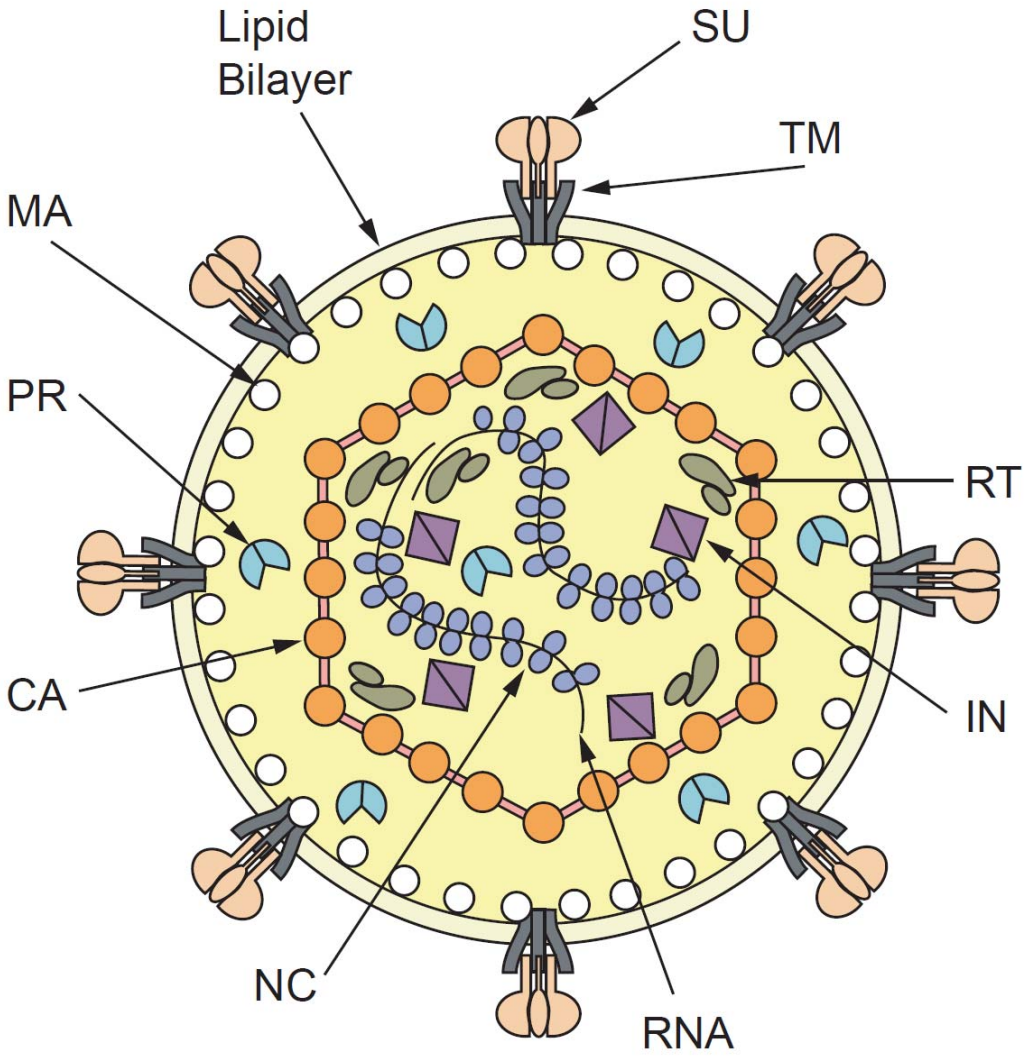


Figure 1.2. Cartooned structure of a retrovirus particle.
Details are provided in the text. Depicted is the overall organization of a typical mature retroviral particle. The dimerized RNA genome copies (shown as curvy lines) are encapsidated within the structural core. Arrows are used to indicate the structural and enzymatic components of the particle. Abbreviations are as follows: From Env: SU, surface unit; TM, transmembrane domain. From Pol: RT, reverse transcriptase; IN, integrase. From Gag: NC, nucleocapsid; CA, capsid; MA, matrix. Pr, protease. Figure is from Voisset, C. and Andrawiss, M. 2000 [28], and Vogt, P.K. 1997. In *Retroviruses*, edited by Coffin, J.M. *et al*. [29], with permissions.

Classification of exogenous and endogenous retroviruses

Historically, retroviruses were classified according to morphology as observed by electron micrography, with four common types referred to as A, B, C, and D [30]. The common morphologies and representatives of each type are briefly described below (referenced to virus descriptions from the International Committee on the Taxonomy of Viruses database; ICTVdb and [31]).

According to this classification system, "A-type" describes intracellular particles or intracisternal particles (IAPs), that bud into the cellular intracisternal space. The IAPs represent the only non-enveloped retroviral type, although a subset of mouse IAPs contain an *env*-like reading frame, and are distinguished as IAPE. "B-type" describes enveloped, extracellular particles that have a condensed acentric core and visible envelope spikes. The representative B-type is Mason Pfizer monkey virus (MPMV). "C-type" refers to enveloped, extracellular particles with morphology much like B-type, but with a rounded central core and less visible envelope spikes. Classic representatives are mouse leukemia virus (MLV), avian leukosis virus (ALV), and also human T-cell leukemia virus (HTLV). *Spumaretroviruses* resemble C-type morphology, however exhibit a less condensed core and more visible envelope spikes. "D-type" describes enveloped, extracellular particles that are slightly larger in size (observed up to 120 nm), and atypical core, and less prominent envelope spikes. D-type retroviruses are represented by the mouse mammary tumor viruses (MMTV).

The modern taxonomy is based on phylogenetic classification, which has led to the re-organization of the morphological categories described above. Detailed phylogenies based on highly conserved sequences, for example within the conserved RT

motif of *pol*, demonstrate that distinct genera are observed within the *Retroviridae* family: *Alpha-*, *Beta-*, *Delta-*, *Gamma-*, *Epsilon-*, *Spuma-*, and *Lentiviruses* [32]. Although in most cases ERVs are no longer replicating, long ago inactivated by mutations, they essentially represent retroviral 'fossils', bearing marked similarities to their exogenous counterparts. Thus, ERVs are also classified according to the same parameters as exogenous retroviruses (XRVs) and offer some power for resolving "missing links" in phylogenetic inference of retroviruses [32, 33].

ERVs have traditionally been categorized according to class (I, II, or III, as mentioned briefly above). Phylogenetic analysis of representatives from each class, along with described XRVs, has established each class groups loosely into one or more of the seven genera within *Retroviridae*. Class I ERVs are most close to the *Gammaretroviruses*, or are '*Gamma*-like' (C-type), class II are *Beta*- and *Alpha*-like (B- and D-types), and class III are *Spuma*-like (most reminiscent to the C-types) (Table 1.1) [22, 32, 33]. However, ERVs have been detected for all genera with the exception of the *Deltaretroviruses* [33]. Most recent were the discoveries of ERVs belonging to the *Lentiviruses*, with the identification of RELIK (for rabbit endogenous lentivirus type K) in the European rabbit [34, 35], and pSIV (for prosimian immunodeficiency virus) in the gray mouse and fat-tailed dwarf lemurs (pSIVgml and pSIVftd) [35, 36]. Thus, categorization of ERVs by class may be complicated as additional elements are identified in the future.

Table 1.1. Characteristics and representatives of the *Retroviridae*.

Genus	Representative XRV	Type[a]	ERV class[b,c]	Representative ERV[b,c]
Alpharetrovirus	ALV	C - avian	II	ALV
Betaretrovirus	MMTV, JSRV	B and D	II	MMTV, also HML-2
Gammaretrovirus	MLV	C - mammalian	I	MLV, GaLV
Deltaretrovirus	HTLV, BLV			
Epsilonretrovirus	WDSV		I	WDSV-like? [d]
Lentiretrovirus	HIV-1			RELIK, pSIV
Spumaretrovirus	HFV, PFV	C	III	HERV-L, HERV-S? [d]

[a] According to International Committee on the Taxonomy of Viruses database; ICTVdb.

[b,] From Gifford, R. and Tristem, M. 2003. [37].

[c] From Bannert, N. and Kurth, R. 2004. [38].

[d] The *Epsilon*- and *Spumaretroviruses* have no described closely related ERVs; ERVs indicated above for these specific Genera are distantly related to the representative XRVs for each group, and cluster weakly in phylogenetic analyses, owing to distant relation.

Another popular method for categorizing ERVs, especially those associated with the more well-documented genera and with reference to HERV groups, has been the letter code for the tRNA complementary to the primer binding site (PBS) used in reverse transcription [33]. For example members of HERV-W have a PBS complimentary to the 3' end of a tryptophan tRNA, and members of HERV-K a lysine tRNA. Within each of such HERV groups, individual elements are sometimes referred to arbitrarily by number, for example HERV-K113 and –K115. Superficially, this methodology is useful for referring to HERVs belonging to distinct phylogenetic HERV groups. Problems with this classification may be encountered for HERVs with high sequence similarity and phylogenetic support, but with PBS sequences more similar to unexpected tRNAs. Furthermore, the PBS sites may no longer be intact for some ERVs as they have been mutated, or even deleted, while in the host genome [37]. This nomenclature has also led

to the tendency to refer to each of these groups as 'families', although technically inaccurate, as 'family' refers to the *Retroviridae* [33].

The varied attempts to maintain such classification nomenclature for the past few decades have led to confusion and disagreement with the naming of ERV groups. Also confusing has been the assignment for individual elements within each representative group, as in some reports each is referenced by an assigned number, while in other reports the same locus may have been given a unique name (also see [33, 37]). A few reports have envisioned a modified classification system reliant on a naming reminiscent of eukaryotic transposons, in which ERV names conform to those of other repetitive elements [33, 39]. Under this system, ERV 'groups' are assigned a name (i.e. ERV-K) from which individual elements are successively numbered (i.e. ERV-K1 to ERV-K_). Whether this suggested nomenclature will be adopted is to be foreseen, however the research community would benefit from a common method of classification.

Retroviral genome organization and gene products

Overview. With few exceptions, retroviral RNA genomes are synthesized and processed exclusively by the transcriptional machinery of the host cell. Each transcript is modified by host enzymes with the additions of a 5' G^7cap and 3' poly(A) tail, and because of this resembles a eukaryotic mRNA [40]. However, and in contrast to other (+)sense RNA viruses, the retroviral genome does not behave as an mRNA immediately following host cell entry and requires first that the RNA be reverse transcribed into a double-stranded DNA copy (viral DNA, or vDNA) that is integrated into the genomic

DNA of the host. It is from the expression of the integrated retroviral DNA, or provirus, that the viral-encoded proteins required for replication are synthesized [41].

The general retroviral RNA genome structure and integrated DNA forms are detailed in Figure 1.3. Preserved between the RNA and DNA forms are the coding regions for *gag*, *pro*, *pol*, and *env*, which are expressed as precursor polyproteins and processed after viral assembly. Retroviruses that encode these essential genes alone for replication are referred to as simple retroviruses. Complex retroviruses encode additional reading frames for expression of accessory proteins, which are generally involved in regulating the transcription of the provirus, the splicing and export of viral mRNAs, or which antagonize inhibitory proteins expressed by the host cell [31, 40].

Non-coding elements. Aside from the added transcriptional modifications by the host, the RNA genome is flanked by direct repeats (R), to which the non-coding unique regions U5 and U3 are immediately internal. The U3, R, and U5 elements are joined during reverse transcription, in that order, to generate the LTRs at either end of the provirus. The LTRs possess critical regulatory elements for its transcription. Just downstream of the U5 region is a non-translated leader (L) region, which contains the tRNA-specific PBS required for the initiation of reverse transcription, and packaging signal (Ψ) essential for dimerization of progeny RNAs and their assembly into particles. At the genomic RNA 3' end and upstream of the U3 region is a polypurine tract (PPT) of ~10 nt required for (+)strand synthesis during reverse transcription [31].

Retroviral genes. The *gag* ORF encodes a precursor protein from which the retroviral MA, CA, and NC structural proteins, that form the inner part of the virion, are derived (Figure 1.2). A Gag precursor contains several determinants that function in virus

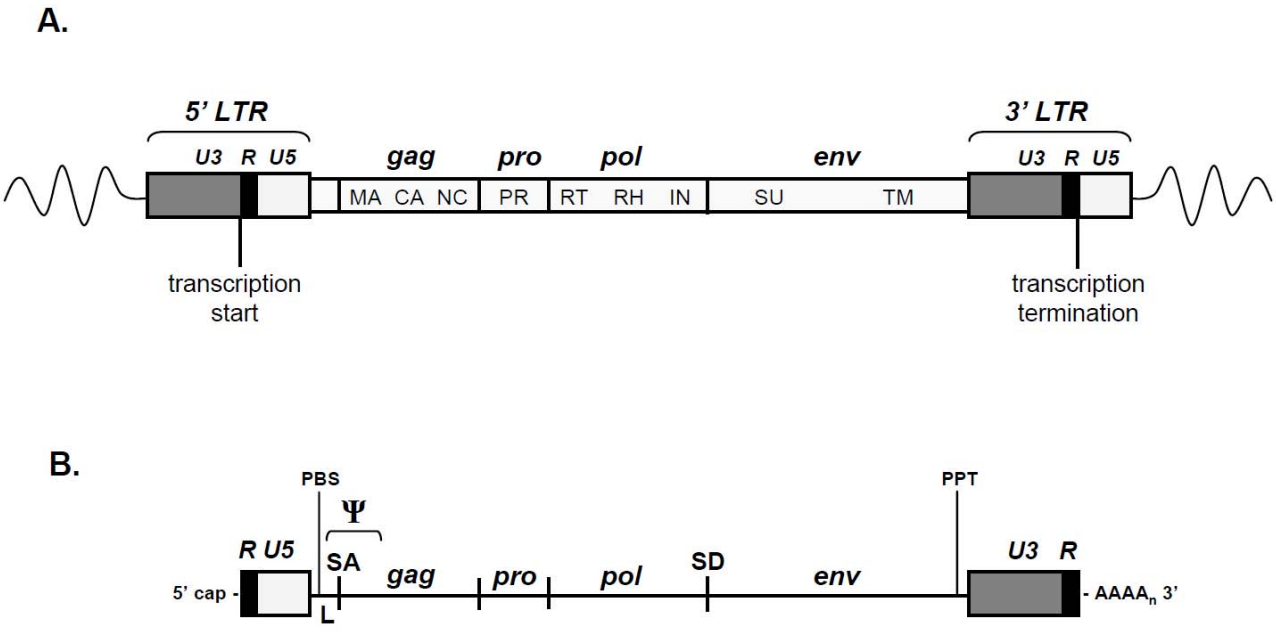


Figure 1.3. Cartoon representation of the structures of an integrated provirus and unintegrated RNA genome.

A. A generalized provirus is shown as integrated into host DNA (indicated by wavy lines). The essential coding genes *gag*, *pro*, *pol*, and *env* are labeled according to relative positions within the provirus. The structural order of the coding regions is shared in all retroviruses. Individual protein subunits and LTRs are abbreviated as in Figure 1.2. The U3-R and R-U5 boundaries are denoted as the sites for the start and stop of transcription of the provirus.

B. The corresponding unintegrated (+)RNA, labeled as in (A). The RNA genome is modified by the host transcriptional machinery to include a m^7G cap at the 5' end, and a 3' poly(A) tail. Splice acceptor (SA) and splice donor (SD) sites for the splicing of the *env* mRNA are indicated. Also shown are the primer binding site (PBS); polypurine tract (PPT); untranslated leader region (L); and packaging signal (Ψ).

assembly including Gag-lipid, Gag-Gag, and Gag-RNA interactions, in addition to others [42, 43]. These functions are facilitated, in part, through the conserved genetic organization of *gag*, whose orientation ensures the correct positioning of the proteins involved in virus assembly [43]. In most retroviruses the MA is co-translationally modified by myristoylation at its N-terminus, and this fatty acid, in addition to a stretch of highly basic residues near the N-terminus, is essential for the efficient targeting and association of Gag to the cellular membrane [42, 44]. At the opposite end of Gag precursor is NC, which selectively interacts with Ψ signals of progeny viral RNAs through its conserved CCHC zinc-finger motif (consensus is $CX_2CX_4HX_4C$) [44], and is responsible for their encapsidation during assembly. The CCHC:Ψ interaction is generally restricted to closely related retroviruses, thus limiting cross-packaging with distantly-related viruses, for example during co-infection or from expression of endogenous proviruses [37].

For all retroviruses, the *pro* (protease) and *pol* (polymerase) ORFs are located downstream of *gag* (Figure 1.3) and encode components of the core precursor polyprotein Gag-Pro-Pol, the synthesis for which differs mechanistically between retroviral types. The *pro* ORF is translated following bypass of the *gag* termination codon either by inframe read-through translation (found in the *Gammaretroviruses* such as MLV), or through a frameshift of the ribosome (-1 base) resulting in continued translation into the overlapping *pro* ORF (found in most other retroviruses) [41]. Similarly, translation of the *pol* ORF differs by virus type and may continue inframe (for example in *Lentiviruses* such as HIV), but may also exist within a third reading frame and

translated following a second frameshift event (observed in the *Betaretroviruses*, for example MMTV) [41].

The primary bypass in Gag termination has a frequency from about 5% to 20% and is recapitulated in the capsid ratios of Gag and Gag-Pro-Pol products, and in the mature and/or active forms of each protein [41, 42]. Pro is responsible for the cleavage and processing of the Gag and Pol precursor proteins, at a limited number of sites, into their mature and active subunits [41]. The cleavage of Gag results in MA, CA, and NC, but can also generate shorter products required for formation of infectious particles in many retroviruses [31]. Cleavage of Pol generates the active forms of RT and IN. RT catalyzes the conversion of the retroviral RNA genome into a cDNA copy through activities of reverse transcription (RNA-dependent DNA polymerase), DNA synthesis (DNA-dependent DNA polymerase) and RNaseH (degradation of RNA when hybridized with DNA in a heteroduplex). IN catalyzes the concerted integration of the linear vDNA product into a nuclear chromosome of the host cell. The active site within this particular enzyme is highly conserved, and is comprised of three invariant residues, D-$X_{(39-58)}$-D-X_{35}-E, known as the $DD_{35}E$ motif, in all retroviral and retrotransposon integrases [45]. Because Pol-mediated cleavage occurs during maturation of the viral particle, both enzymes are transported in their active forms in the virion to the next target cell.

The *env* mRNA (Figure 1.3) is generated through subgenomic splicing from the major splice donor (SD) site located downstream of the PBS in primary proviral transcripts, and encodes an N-terminal signal peptide which targets its translation to the rough endoplasmic reticulum (ER) [44]. Env is translated as a heavily glycosylated precursor protein into the lumen of the ER, and later processed while oligomerized into

the SU and TM subunits during trafficking through the Golgi network. The cleavage is mediated by a host-encoded protease (usually furin or furin-like) and is at a highly conserved motif: K/R-X-K/R-R [44]. Env proteins reaching the cell surface of the infected cell are included in the lipid bilayer envelope of budding virions [44]. During cellular infection, SU and TM each have distinct functions in binding receptor(s) on the target cell, and in membrane fusion, respectively. The Env-receptor interaction mediated by SU is highly specific and determines the host range of the retrovirus [46].

The retroviral replication cycle

Entry and uncoating. A retroviral infection cycle (Figure 1.4) begins with an interaction between the viral SU and a specific receptor (and sometimes co-receptor) on the host cell surface. The receptor-bound complex induces dramatic conformational rearrangements in the SU-TM subunits, enabling the fusion peptide, located at the N-terminus of TM, to insert directly into the host cell membrane, thus triggering fusion of the viral and host cell lipid bilayers. For some retroviruses, the binding of SU and receptor triggers endocytosis of the virion, after which fusion occurs with the endocytic membrane intracellularly [46]. The capsid is released into the host cytoplasm where 'uncoating' occurs, during which the virion core capsid proteins are dissociated and/or reorganized into a nucleoprotein complex (also referred to as the reverse transcription complex, or RTC) in which reverse transcription takes place [41, 47].

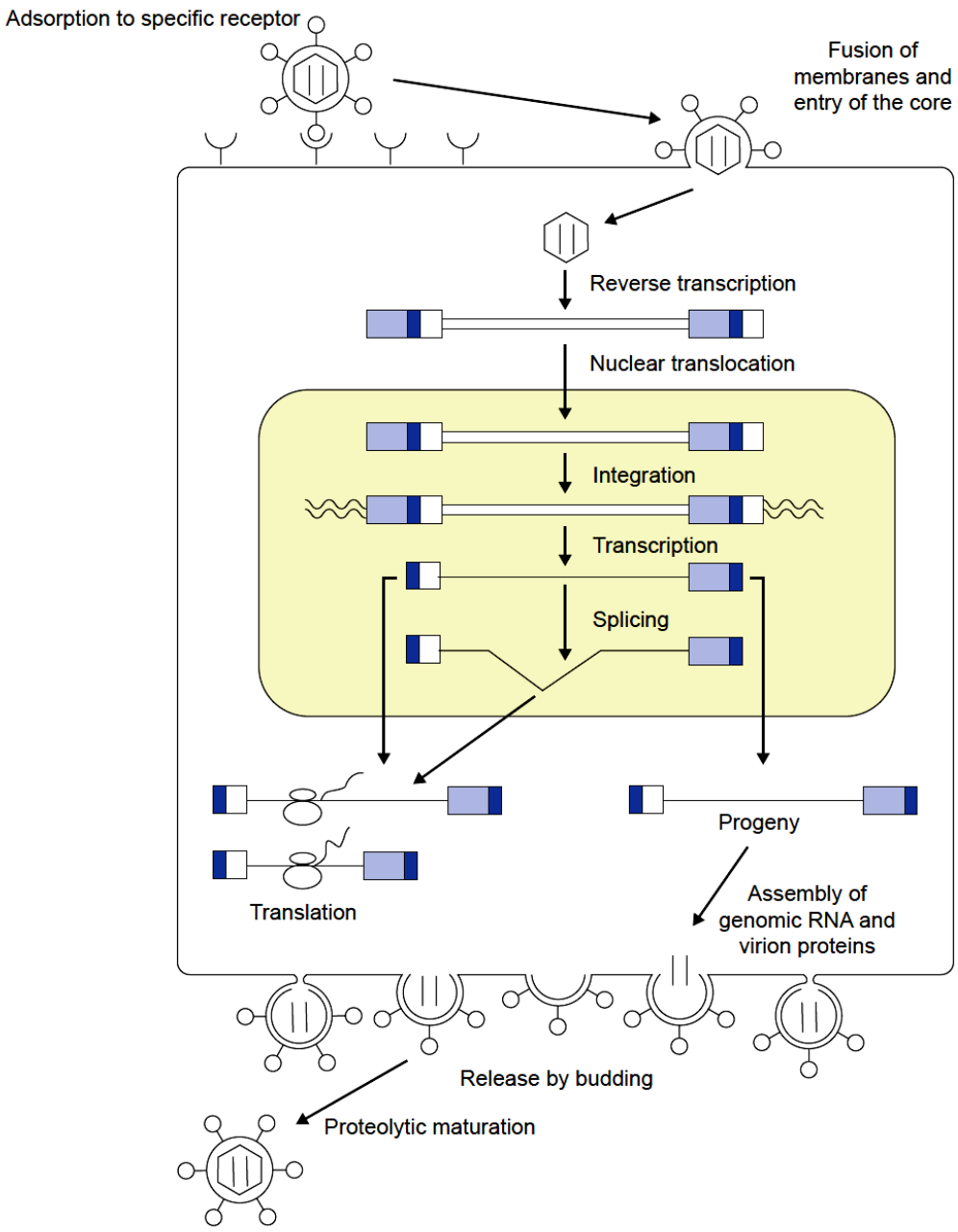


Figure 1.4. Diagram of the retroviral replication cycle.
The steps of a typical infection cycle are shown; details are further provided in the text. The virion core is depicted within the lipid bilayer (outer circle), with Env subunits displayed on outer surface. The two RNA genomes are shown as straight lines within the core. The unintegrated vDNA and integrated provirus are represented as in Figure 1.3. U3 is light blue; R is dark blue; U5 is white. Ribosomes are depicted in translation of the viral mRNAs. Tan shading indicates the nucleus of the host cell. Figure is from Voisset, C. and Andrawiss, M. 2000 [28], and Vogt, P.K. 1997. In *Retroviruses*, edited by Coffin, J.M. *et al*. [29], with permissions.

Reverse transcription. Reverse transcription is one of the defining features of retroviral replication and is diagrammed in Figure 1.5 [41]. The ability of RT to catalyze the reverse transcription of the viral RNA genome into its DNA replica relies on its intrinsic activities of RT and RNaseH nuclease. Other non-coding sequences, including the R, U3, U5, PBS, and PPT regions have significant roles in this reaction and function in the generation of the LTRs and for sites of primer binding and elongation.

The reaction is outlined as follows. Reverse transcription is primed from the 3' end of the host tRNA hybridized at the PBS; this site differs between retroviruses, and determines the tRNA used in priming the reaction. Briefly, the R-U3 region located just upstream of the PBS is the first part of the RNA genome to be reverse transcribed. This reaction initiates synthesis of the vDNA minus strand to the 5' edge of the genomic RNA, resulting in the formation of the minus strand strong-stop DNA. As reverse transcription progresses, the RNaseH activity of RT removes the template RNA from the resulting RNA:DNA hybrid, releasing the newly synthesized product. By virtue of complementary R regions, the minus strand DNA is transferred to the 3' edge of the genomic RNA in the first of two strand transfer events. Elongation by RT continues through the U3 to the opposite end (the 5' edge of the PBS), with the concomitant degradation of the complementary RNA strand by RNaseH.

Near the 3' end of the genomic RNA, upstream of the U3 region, is an ~11 base tract of purines (PPT). This short stretch is resistant to RNaseH degradation, and the template RNA remaining at the site serves as the primer for plus strand synthesis. From the 3' edge of the remaining genomic RNA, RT elongates using the vDNA as a template through the U5 and to the edge of tRNA remaining on the minus strand, with RNAseH

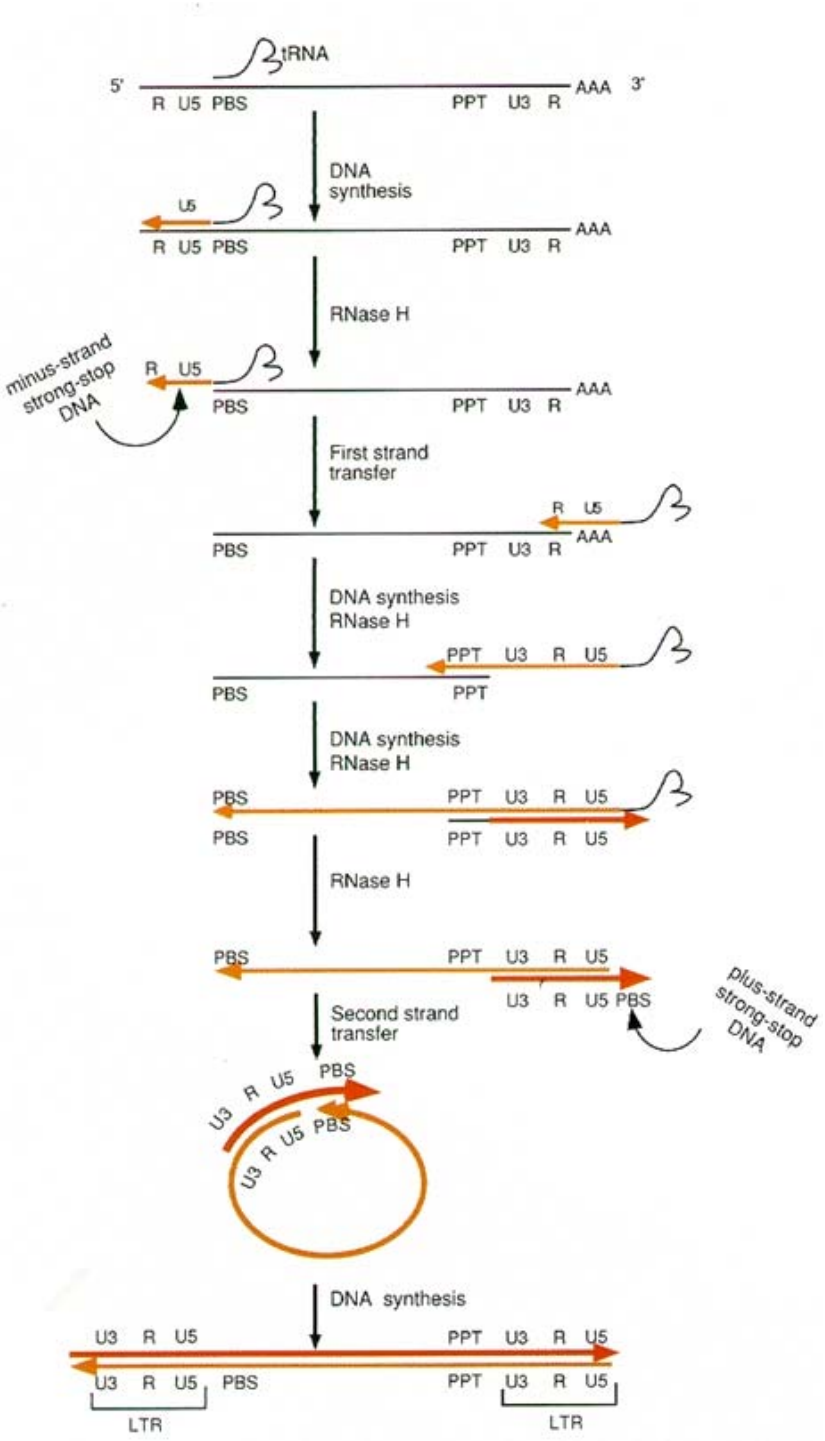


Figure 1.5. Diagram of the steps in reverse transcription.
Details are provided in text. All abbreviations are as in previous figures; see Figure 1.3. From Telesnitsky, *et al*., 1997. In *Retroviruses*, edited by Coffin, J.M. *et al*., with permission.

removing the original tRNA. This results in the synthesis of the first portion of the plus strand, referred to as the plus strand strong stop. A second strand transfer occurs, in similar fashion to the transfer of the first strand, however mediated by complementary PBS sites. From the plus strand PBS, reverse transcription continues to the opposite end of the already formed minus strand to complete plus strand synthesis, and resulting in the generation of the identical LTRs at either end of the vDNA [41].

Two notable properties of RT are manifested through reverse transcription. First, RT is an error-prone enzyme, and exhibits a high rate of misincorporation averaging about 1 base change per 10kb per replication cycle due to the lack of a proof-reading 3'→5' exonuclease activity and mismatch repair [41]. Second, RT has a relatively low template affinity and undergoes 'template switching', or 'jumping', between RNA strands during reverse transcription, which can lead to the generation of novel recombinant sequences in the presence of non-identical co-packaged parental RNAs, or which can function to overcome defects that may be present in one of the RNA copies [48]. In combination with the RT error rate, these features allow for the rapid evolution of an infecting retroviral population in the presence of selective pressures such as the host immune response, or in case of retroviral inhibitors [49].

Nuclear transport. Following synthesis of the vDNA, the reverse transcription complex is converted into the pre-integration complex (PIC) in preparation for transport to the nucleus and the integration of the double-stranded vDNA into the host genome [50]. The PIC, which includes at least the viral IN and vDNA, is sufficient for *in vitro* integration of vDNA into a target DNA, however *in vivo* integration requires that the PIC cross the nuclear membrane. Different retroviruses have been shown to use

different strategies to gain access to the host nuclear DNA [51]. In general, the simple retroviruses, such as MLV, are dependent on progression of the cell cycle, and replication is restricted to actively dividing cells. However some simple retroviruses are able to infect some types of non-dividing cells, though with low efficiency [50]. Complex retroviruses, such as HIV-1, have evolved alternative strategies to promote the active transport of the PIC across the nuclear membrane, for example via interactions with host factors, thus allowing for integration in non-dividing cells [52].

Integration. IN catalyzes the integration of the retroviral DNA into the nuclear genome of the host, and results in the formation of a provirus. Integration proceeds in three steps that are common to all retroviruses and are briefly described as follows (Figure 1.6). Within the PIC, IN catalyzes the removal of the 3' terminal dinucleotide from either end of the newly synthesized linear vDNA in a reaction known as 3' end processing. Following nuclear transport, IN then catalyzes the concerted reactions of cleavage and ligation during strand transfer, in which the processed 3' vDNA ends are joined to the 5' ends of nuclear target DNA at a staggered cleavage site of 4-6 bp. Lastly, cellular DNA repair enzymes fill the gaps flanking the viral DNA, resulting in the short, 4-6 bp target site duplications which are characteristic of all retroviruses [51].

A mechanistic perspective of how these domains interact with vDNA and target DNA (tDNA) during integration has only recently been demonstrated in profound structural studies using the prototype foamy virus (PFV) [53, 54] and are noteworthy. The high-resolution crystal structures of the PFV intasome, or the minimal integration nucleoprotein complex consisting of IN, vDNA, and tDNA, revealed an unexpected and novel organization which differed markedly from existing models (Figure 1.7) [55].

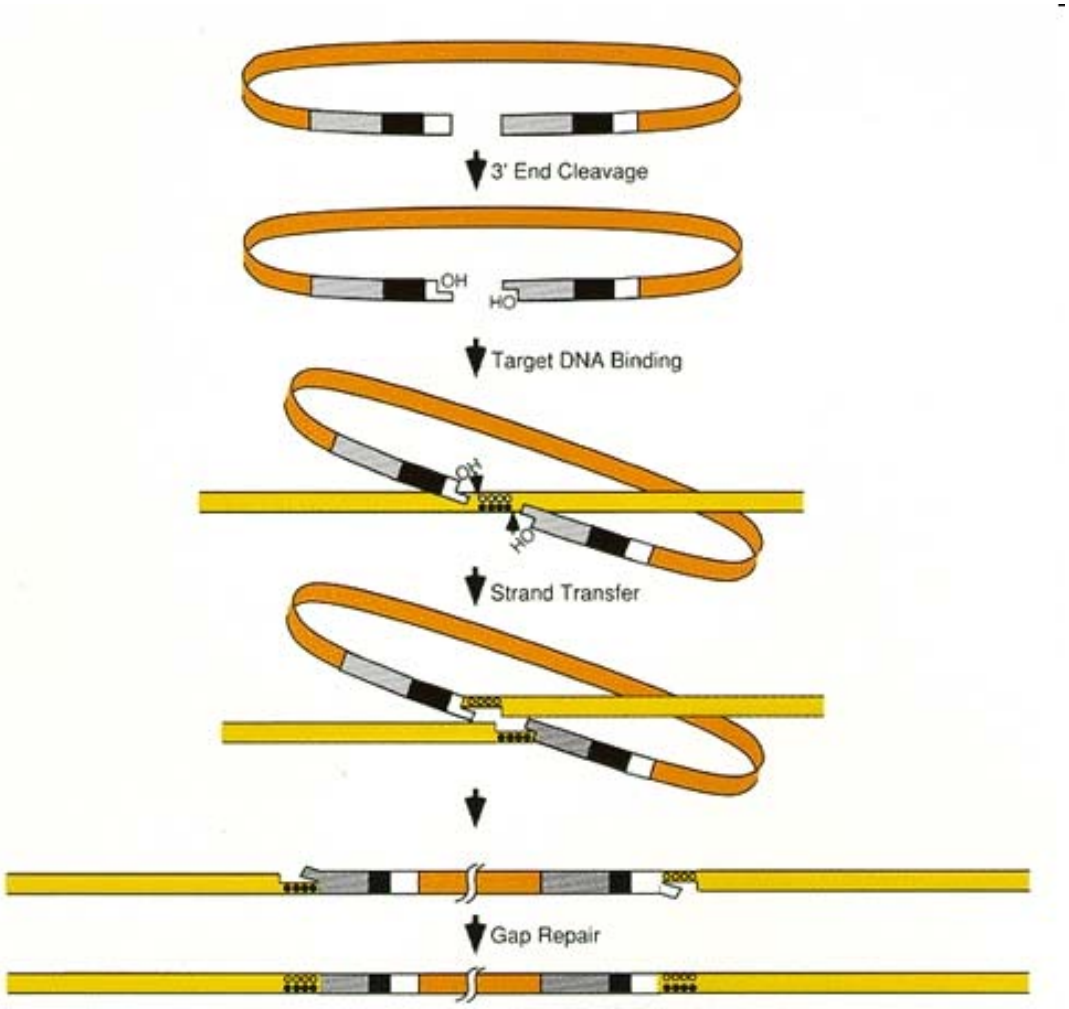


Figure 1.6. Diagram of the steps of integration.
The steps in retroviral integration are shown: 1. removal of the 3' dinucleotides in 3' end processing, 2. the concerted cleavage-ligation reaction in strand transfer, and 3. covalent joining of the vDNA to the host sequence. Further details are provided in the text. LTRs are indicated at the vDNA ends by shading as in previous figures. Viral DNA is shown in orange, and the host target DNA in yellow. Open circles within the target DNA are to indicate the plus strand, and filled circles the minus strand in the host genome. From Brown P.O., 1997. In *Retroviruses*, edited by Coffin, J.M. *et al*., with permission.

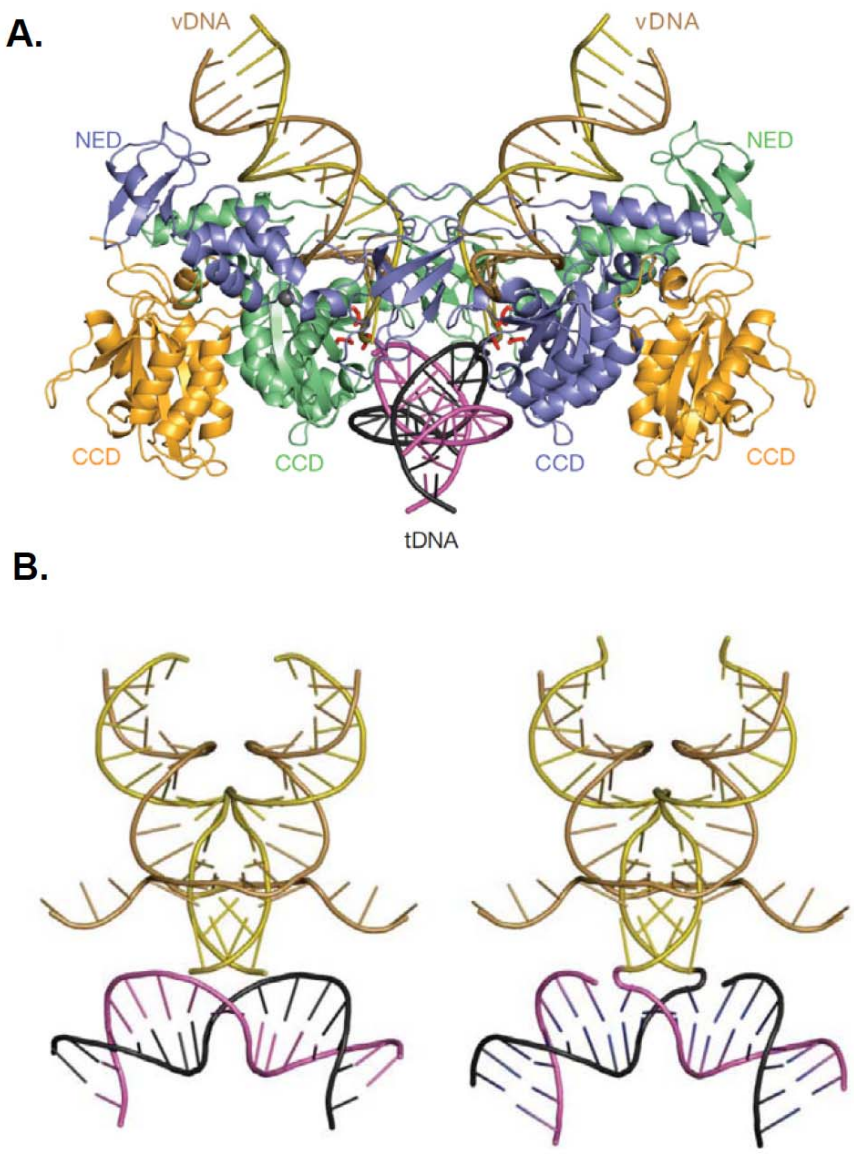


Figure 1.7. Molecular view showing the integration of retroviral DNA.
Shown is the crystal structure of the PFV intasome. Labeling is as follows: vDNA is in yellow and gold; tDNA is pink and black. IN is shown in ribbon form with individual domains of the IN tetramer colored and labeled: CCD, catalytic core domain; NED, N-terminal domain. A. The crystal structure shows the interactions between the IN tetramer, vDNA, and tDNA. The inner two CCD domains form the active site groove, into which the tDNA is precisely positioned. The outer CCD and NED domains are thought to stabilize the interaction. The dinucleotides of the vDNA are shown near either tDNA strand in red. B. Structure of tDNA and vDNA prior to and following strand transfer. Figure from Maertens G.N., Hare S., and Cherepanov P. 2010. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* 468(7321):326-9., with permission.

While it had been previously shown that IN was associated with the vDNA as a tetramer, it was revealed that only two of the four available active sites are in contact with the recessed 3' ends of the vDNA, and the remaining IN monomers function to stabilize the complex in a highly elaborate system of IN:IN and IN:DNA interactions [53]. The resulting conformation induces a severe bend of 55° in the tDNA and at the enzyme's active site, precisely directing the sites of target cleavage (which, in the case of PFV, generates a gap of 4 bp) [54]. The mechanism and stoichiometry are consistent with previous biochemical studies, and provides support that certain structural features in DNA, such as distortion induced by nucleosome association, act to improve the efficiency of target site integration (also reviewed in [56]).

Except for the absence of two bases at either end of the vDNA, the provirus is identical to its unintegrated form, with invariable terminal dinucleotides of 5' TG/CA 3' that delineate its junction site within the host genome. Also characteristic of the integrated provirus are short and sometimes imperfect inverted repeats, either of which is situated just internal to the 5' and 3' terminal dinucleotides, and thought to provide recognition sites for the viral IN (Figure 1.8) [51]. Aside from rare recombinagenic events there is no specific mechanism known for excision of the provirus. The integrated element is permanent and behaves as a cellular gene. This reaction is essential to establishing a productive infection and is the second defining feature of retroviral replication [51].

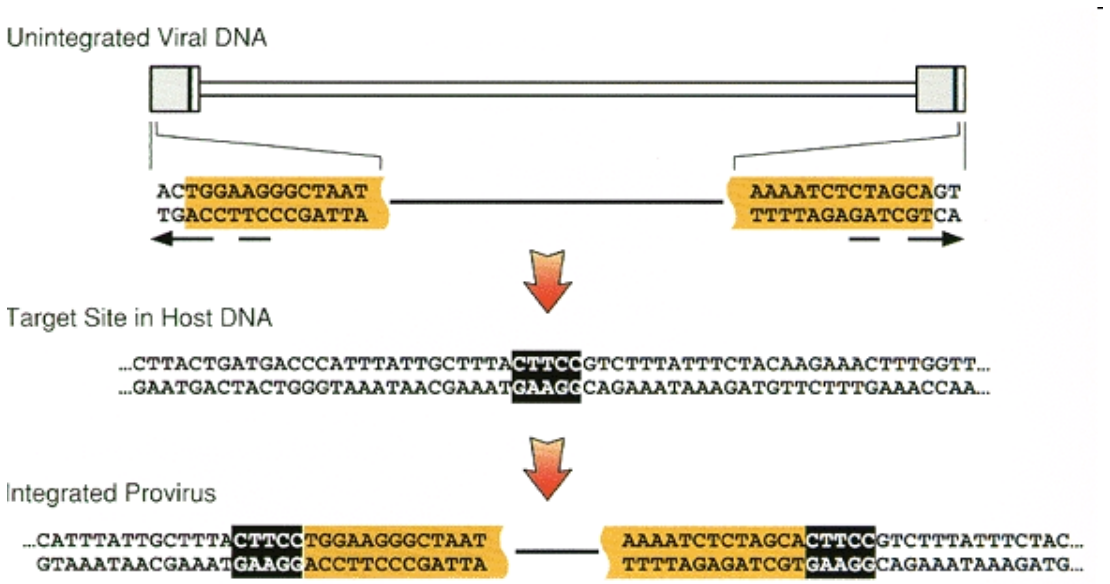


Figure 1.8. Nucleotide features of a proviral integration site.
Shown is a schematic detailing the sequence composition at the edges of the vDNA prior to its integration, and the sequence features in the resulting proviral junction site. Shaded in yellow are the ends of the viral DNA pre and post integration. Black shading represent the target site sequence, and subsequent TSDs flanking the integrated provirus. White letters indicate the TSD sequence.
Upper. Unintegrated vDNA. The broken arrow underlines the imperfect repeats containing the canonical 5'TG/CA3' dinucleotides located two bases from either edge (bordered by the yellow highlighting).
Middle. Integration target site, showing 5bp target in black background.
Lower. Integrated provirus with detailed junction site. The terminal dinucleotides are removed during integration, resulting in the canonical proviral edges. The staggered cut at the target site is filled in by host DNA polymerase, resulting in the identical TSD.
From Brown, P.O., 1997. In *Retroviruses*, edited by Coffin, J.M. *et al*., with permission.

Proviral expression. The LTRs contain regulatory elements recognized by cellular transcriptional machinery for the synthesis and processing of viral mRNAs and thus promoting expression of the provirus. Situated within the U3 region, near the 5' edge and ~75 bp in length, is the enhancer core region that contains binding sites for varied cellular transcription factors. Downstream of the enhancer region, near the start of the R is the TATA box signaling the start of transcription, and transcription proceed such that it delineates the U3-R edge of the LTR. Located within the R region is the poly-adenylation signal (AATAAA) for transcription termination, and also mRNA transportation [40, 57]. Similarly to the start of transcription marking the U3-R edge, the site of transcription termination delineates the R-U5 edge (Figure 1.3). The transcripts are m^7–capped and polyadenylated as cellular transcripts, a subset of which are spliced to generate mRNAs for *env* and/or accessory proteins [40]. The viral mRNAs are transported to the cytoplasm where they are translated on free ribosomes (Gag and Gag-Pro-Pol) or rough ER (Env), and the proteins subsequently localized to the inner or outer cell membrane, respectively [44].

Assembly and exit. Gag associates with the inner membrane via MA, many times aided by N-terminal modifications such myristoylation, as mentioned above, and adopts a rod-like conformation perpendicular to the host membrane. This orientation thus facilitates the incorporation of the two progeny RNAs via interaction with NC [44]. The virions bud from the cell surface enveloped in a lipid bilayer as premature particles. During maturation, Pro is responsible for cleavage of the Gag and Pol products, at a limited number of sites, during budding from the host cell membrane or shortly thereafter [41, 58]. This results in a series of conformational changes within the virion, visible as a

change in the electron-lucent particle to the 'condensed core' and ultimately results in the mature and infectious virion [44].

Endogenous retroviruses

Formation and fate of an endogenous retrovirus

Retroviruses usually infect and replicate within somatic cells but occasionally will gain access to, and infect germ line tissue. Provided the host cell (and its progenitors) survive, and that the retroviral integration has no immediate detrimental effects, the provirus will be transmitted to the host's offspring as stable genetic material in a Mendelian pattern of inheritance [59, 60]. Proviruses that do become permanent residents of the host's genome are referred to as endogenous retroviruses (ERVs), and are said to be transmitted vertically to contrast from the horizontal transmittance of exogenous retroviruses (XRVs) (Figure 1.9). ERVs have been restricted to the vertebrates, and with the exception of the most basal class represented by the 'jawless fish' (Class: Agnatha; hagfish and lampreys) [61], have been found in every vertebrate class examined, and have contributed to substantial portions of their respective genomes [62].

Retroviral elements are permanent fixtures of the host genome once integrated, without any known mechanism of directed excision [51]. However, there are rare circumstances in which an ERV may be 'removed' from the host's genome. For example, allelic conversion may recover the pre-integration site through nonreciprocal recombination that essentially transfers the empty allelic sequence to the ERV-containing allele [63], or in an ectopic recombination, which involves an atypical (or inter-element)

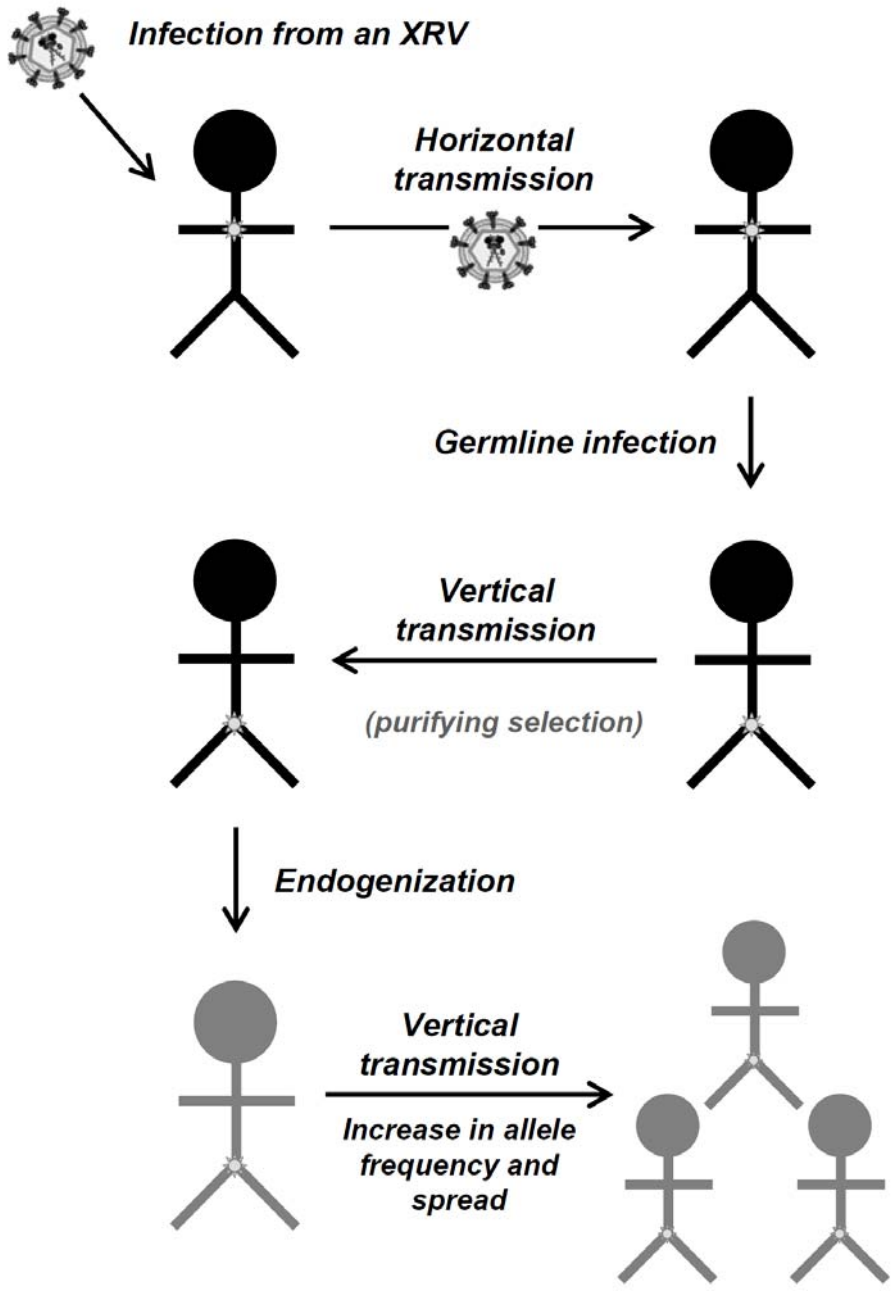


Figure 1.9. Diagrammed steps in the formation of an endogenous retrovirus.
The process begins with the horizontal infection of an exogenous retrovirus (XRV). Occasionally, an infectious particle will gain access to and infect a germ cell or germ tissues (germline infection). If the provirus in the germ cell and/or resulting offspring is not immediately deleterious, it will be vertically transmitted to offspring as endogenous retroviruses. Germline integrations in offspring that are not removed due to purifying selection have the increased likelihood to have continued spread through the population by successive vertical transmission, and eventual fixation.

recombination event between highly similar DNA sequence stretches or chromosomal regions [64, 65]. The likelihood of ERV loss as a result of these types of recombination becomes negligent with increased fixation of the ERV and loss of the insert-free allele from the population [38].

The probability for an ERV to increase in frequency within the population is dependent on several factors, such as the cost/benefit effects to the host, in addition to stochastic effects such as bottlenecks or during bursts of population growth or expansion [38]. Unless having already been inactivated at the time of insertion, ERVs may briefly continue to replicate, the time course of which will depend on effects in fitness to the host. ERVs capable of replication that leads to new germline integrations are known as 'founder' proviruses, sometimes detected as a single clade by tree topology, depending on the divergence between new ERV formations [12]. The replication capacity of such founders effectively decreases as ERVs are inactivated, either from recombination generating a solo LTR or accumulated mutations within the viral coding sequence, as in most cases there is little pressure for the host to maintain sequence integrity of the ERV [12, 37]. Though rare in comparison, inactivation may be slowed dramatically from selective maintenance if the ERV confers benefit(s) to the host (discussed further below in Biological Significance and Impact).

ERVs provide a useful sequence source for evolutionary studies, as many still bear structural similarities to actively circulating XRVs despite the tens of millions of years they have sometimes resided within the host genome [66]. As viral 'fossils' in vertebrate genomic DNA, ERVs represent a unique source for studying virus evolution over many millions of years, given their relative ease of sampling, range of database

sources and genome DNA samples. Characterizing ERVs has power in allowing for the prediction of retroviral phylogeny and classification, which can be problematic when using actively circulating viruses given their high rate of evolution.

Human endogenous retroviruses

Retroviruses have been colonizing the primate genome for at least ~25 million years, as evident from shared loci with Old World monkeys (Figure 1.10) [38, 65, 67]. HERVs have been traditionally divided into 'families', which are referred to throughout this document as 'groups'. Proviruses representing each group were originally identified and characterized by homology to exogenous retroviruses, largely based on PCR analysis and DNA hybridizations, and more recently have been identified through efforts in computational data mining and analyses. With such studies, the total numbers of elements, and their representative groups, have expanded considerably in the few decades since the discovery of HERVs.

The most current sequencing builds of the human genome have shown that greater than 8% of our DNA is of retroviral origin, represented by some ~440,000 elements, for a total genome representation of roughly 225 Mb of DNA [2]. HERVs are represented by as many as 31 phylogenetically distinct groups [22, 37, 68], each having arisen from a single [cross-species] germline infection [68]. However, detailed phylogenetic analysis of the known HERV groups also reveals a number of monophyletic branches, indicating 'missing', or not yet identified ERVs; in the absence of such elements the relationships between other HERVs cannot be resolved. The identification

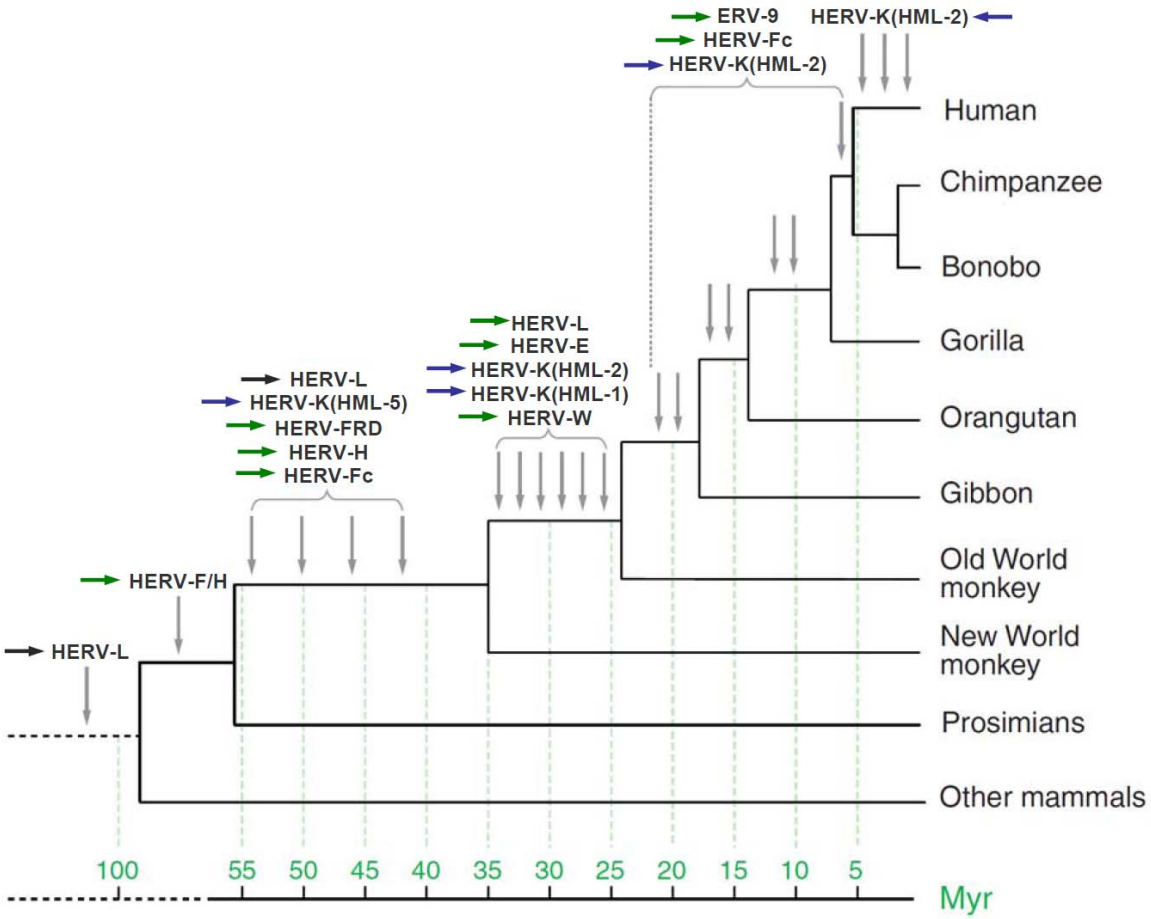


Figure 1.10. Approximate germline integration times for HERV groups. Selected HERV groups are shown. Arrows are meant to indicate the HERV Class for each indicated group: Dark grey, Class III; Green, Class I; Blue, Class II. Estimated divergence times are shown in millions of years, and were from [69]. Figure was adapted with permission from Bannert, N. and Kurth, R. 2006. The evolutionary dynamics of human endogenous retrovirus families. *Annu. Rev. Genomics. Hum. Genet.* **7**:149-73.[38].

of additional ERVs in other species would likely increase the number of HERV groups in the future.

As mentioned, HERVs are further categorized into one of three classes (I, II, or III) from their phylogenetic analysis of a highly conserved motif within RT (also refer to Figure 1.10) [22]. Classes I and III represent older HERVs and are fixed among humans, present in ~110,000 and 83,000 copies, respectively. The class I HERVs are the largest class, represented by as many as 23 groups in ~112,000 elements and nearly 80 Mb DNA. Class III are comprised of 4 groups in ~83,000 copies and 40 Mb. This class also contains the HERV-L, which include some of the most ancient germline insertions, at nearly ~100 mya and prior to the divergence. The evolutionarily 'younger' class II is also represented by 4 HERV groups, including the most recently active HML-2, and collectively has ~8,000 copies over ~8.5 Mb DNA [2].

Within each group there is considerable variation in copy number, depending on the specific group. For example, the HERV-H are present in >1100 copies, while the HERV-E are present in <200 [70]. Although some variation in copy number may be attributed to the length of time for potential activity, some of these discrepancies can be resolved through inference of copying mechanism, and is discussed further below.

Mechanisms of increase in copy number of HERVs in the germline

The vast majority of elements from described HERV groups formed from germline insertion millions of years ago, and were long ago inactivated of any infectious capacity. Only the recently active HML-2 group is considered to possibly contain infectious copies. The activity of such elements would likely have detrimental affects to the host, an outcome of which is their removal from the population under a model of negative selection. Proviruses that are rendered incompetent upon germline infection are more likely to increase in frequency in the population, and may eventually reach fixation. Therefore, it is not surprising that all described HERVs are defective for replication. As 'fossils' of their infectious counterparts, however, their sequence analysis can provide much inference into their mechanism(s) of proliferation within the germline.

The patterns of nucleotide substitutions within genes (or parts of genes, for example where encoding domains or motifs with specific functions) are many times reflective of the imposed forces of selection on the encoded amino acids at those sites. In general, a relative abundance of nucleotide substitutions that do not affect the coded amino acids (i.e., synonymous changes) are consistent with a pressure to maintain those residues at those sites. Likewise, a comparatively high amount of nucleotide substitutions that result in the coding of different amino acids (i.e., non-synonymous changes), or that introduce stop codons, may indicate a relaxed requirement for the specific coded residues at those sites or the functional maintenance of the encoded protein. Thus, comparing the ratio of non-synonymous to synonymous nucleotide substitutions, or d_N/d_S, may be used to infer past selective pressures for a defined coding region. In most cases, an overall trend of $d_N > d_S$ (or a ratio of $d_N/d_S > 1$) is interpreted as positive selection, $d_N \sim d_S$ (d_N/d_S

~ 1) indicates neutral selection, and $d_N < d_S$ ($d_N/d_S < 1$) indicates negative selection for a defined site. However, more than one amino acid may be maintained at a site depending on the imposed selective forces, for example the pressure to evade detection by the host's immune system, or in the presence of anti-retroviral inhibitors, and could have consequence to the overall d_N/d_S estimates when included in such an analysis. In any case, for HERVs, most evolutionary change that occurred prior to germline integration is preserved in the endogenous element. Therefore, the d_N/d_S for such elements, again with consideration of specific functional domains or motifs, can be used to infer the mode of endogenization.

Infection of the germline

Infection of the germline by replication-competent viruses is considered to have been the primary route of HERV formation. Evidence of infection is inferred through a low ratio of $d_N < d_S$ for the *env* gene ($d_N/d_S <<<1$), as it is required for entry to the host cell. Analysis of the *env* gene in this way has shown that, of the major HERV groups (about 17 in all), most elements have evidence of purifying selection of the *env* gene, with d_N/d_S well below 1 [70, 71]. Generally, this has been the observed trend for the HERV groups with comparatively low copy numbers, such as the HML-2 and the HERV-E, present at ~100 and 200 copies respectively, and with a d_N/d_S near 0.2 for *env*. Few HERV groups exhibit *env* d_N/d_S ratios approaching 1; those groups that do tend to have copy numbers in the hundreds and even thousands, and *env* d_N/d_S ratios of ~0.75 to 0.95 [70]. Alternate copying mechanisms can be inferred for such cases as described below.

Complementation in trans

The HERV-H group is most interesting in its division into major and minor subgroups, respectively represented by elements with large shared deletions (>1,200 copies), and those which are mostly intact (less than 100 copies) [72]. The *env* d_N/d_S between the groups is quite striking, at 0.95 for the major group, and just 0.25 for the smaller intact group; in the case of the former major group, this d_N/d_S value is the closest to =1 of any analyzed HERVs [70]. This pattern can be expected for complementation *in trans*, in which defective elements are co-packaged and subsequently copied through the functions of intact members of the same HERV group, under the conditions that critical *cis*-regulatory sequences remain intact (i.e., the packaging signal ψ, LTRs, PBS, and PPT), and that the essential viral enzymes and structural proteins are supplied from related ERVs in the same cell or from an infecting XRV [12]. As a result, defective elements give the appearance that they have actually been reinfecting. This can also be detected in a phylogeny, for example in a tree representative of the HERV-H reading frames, the defective, largely deleted elements fall into a large monophyletic clade, while the minor intact elements form a smaller clade of their own [73]. Finally, HERVs may also continue to replicate *in trans* with other functional or partially defective expressed ERVs, or from an infecting XRV, which may occur when transcripts from distinct proviruses (but which may contain recognizable packaging signals) are packaged into a single infectious particle [12].

Retrotransposition in cis *and other modes of proliferation*

Other HERVs provide evidence of ERV formation from an additional, though rare, copying mechanism. The HERV-K(HML-3) group exhibits a high *env* d_N/d_S of ~0.7, but interestingly the same analysis of its *pol* gene provides a d_N/d_S of levels comparable to the *env* of reinfecting HERVs, around 0.15 [70]. Such a pattern could be expected for elements able to copy by the mechanism of retrotransposition *in cis*, in which elements replicate themselves, but do so intracellularly. As a result, functional *gag* and *pol* proteins would be required, but the absence of infection would be no longer necessitate a functional *env*. To be sure, the HERV-K(HML-3) also have evidence of complementation *in trans* from a subset of elements [74].

There are other HERV groups that have not yet been subjected to similar analyses, however some are the remnants of ancient infections, and are not associated with *env* genes. Likely, such elements proliferated within the germline in an intracellular mode. A final mechanism, rare and only observed for a subset of the HERV-W group, is a sort-of 'piggyback' copying mechanism akin to that used by *Alu* and SVA elements, as in this case the HERV-W elements have been copied by LINEs, reminiscent in their appearance as processed pseudogenes with poly-A tails (also refer to the Class I non-LTR retroelements in Figure 1.1). For this subset of HERV-W, this has apparently occurred multiple times, evident in a phylogeny of the whole HERV-W group, in which the LINE-copied elements are scattered through the tree. The HERV-W group tree topology still suggests, however, a subset of elements from infection [70].

Overall, the predominant mode for the proliferation of different HERV groups appears to have been through germline infection. However, in comparison to the elements that have had much success increasing copy number via complementation and/or retrotransposition, exogenous infection has not led to the high numbers observed for such other HERV groups. This is likely an outcome of the prediction that most germline integrations from infectious competent elements are detrimental to the host, and without expedient inactivation, are lost in natural selection.

Biological significance and impact of HERVs

The fact that at least 8% of the human genome is of retroviral origin suggests HERVs have been an important factor in the course of our evolution. For instance, repeated germline infections represent the addition of new genetic material as a source of potential coding sequence to the host. HERVs also provide a template for a variety of recombination processes both within and between elements, serving as a means for structural rearrangement and reorganization [75, 76]. Beyond the potential to influence genome structure, HERVs may influence the regulation of host gene expression as promoters or enhancers, through post-transcriptional splicing of host mRNAs, or by introducing alternative polyadenylation sites [75, 77]. Examples are discussed below.

HERV distribution in the genome

A few studies have more closely investigated HERV distributions throughout the human genome. One report found that the majority of the Class II HERVs, including the younger HML-2, were skewed toward genome regions having relatively high GC%

(used as an inference of gene density), in contrast to the more ancient Class I and III HERVs, which follow a general trend toward lower GC% regions [78]. In an extension of these results in the same report, the distributions of Class I, II and III HERVs were observed to vary strikingly in proximity (up to ~30 kb) upstream and downstream of annotated host genes. Most HERVs belonging to Class III, collectively the oldest elements, were most dense within ~20 to >30 kb, and underrepresented within ~10 kb from host genes, in contrast to the Class II elements, which were most dense within ~5 to 20 kb [78]. Class I HERVs, consisting of elements intermediate in age to Classes II and III, were found over a broader range of ~5 to 30 kb [78]. HERVs of any group are rarely found within genes [78, 79], with exceptions including a few HML-2 [80]. Of note, HERVs within host genes tend to have anti-sense orientation, likely from selection against HERVs that negatively affect host gene expression via transcriptional signals encoded within the provirus [78, 79, 81, 82].

The observed pattern for HERVs with proximity to host genes, in combination with the relative time of germline entry for Class I, II, and III HERVs (also refer to Figure 1.10), suggests an evolutionary trend in which, following integration near genes, HERVs are selectively "pushed away" from cellular genes over time. This model proposes integrations close to genes have negative effects to the host and are thus progressively removed [78, 81]. A caveat is in the application to individual elements as a function of age and gene distance, due in part to the reliance on age determination via reference to consensus group sequences [78]. Possibly, the estimated age/gene proximity pattern is reflective of successive integrations of tolerable proviruses, resulting in the more neutral 'shifting' of inactivated older HERVs as new elements are acquired.

Additional support for this pattern has been from studies addressing HML-2 alone. Previous chromosomal mapping studies reported an enrichment of HML-2 LTRs (including those associated with proviruses) within the 19p12–19q13.1 regions of chromosome 19 that encodes an array of zinc finger (*ZNF*) genes [83], and a correlation between HML-2 LTR density and host gene density on chromosome 21 [84]. More recently a genome-wide comparison of conserved HML-2, the vast majority of which were solo LTRs, found that about half of tested elements (76/156) were associated with genes: 24 were located ~5 to 35 kb from the associated gene; 40 were <5kb away, and 12 were within genes [85]. This analysis provides some support for the 'shifting' pattern described above, as the included elements were conserved and the majority were human-specific, the results suggest a concentration of more recently formed elements near genes.

HERV impact on host genome structure

Transposable elements provide a significant source of nucleotide sequences available for various rearrangements throughout the genome, owing to the magnitude in copy numbers making various groups, and their inherently repetitive nature. Retroelements have been given much attention in this context, particularly so for the non-LTR elements, which have been shown to mediate large-scale rearrangements including, but not limited to, deletions, duplications, and inversions [3, 86]. HERVs have also been implicated in such events, however have not been as widely reported as for the *Alu* and L1 elements. Nevertheless, there is well-supported evidence for retroviral-mediated rearrangements that generally fall into two categories: 1. those between LTRs (proviral or solitary) (Figure 1.11), and 2. those involving internal proviral sequence (Figure 1.12).

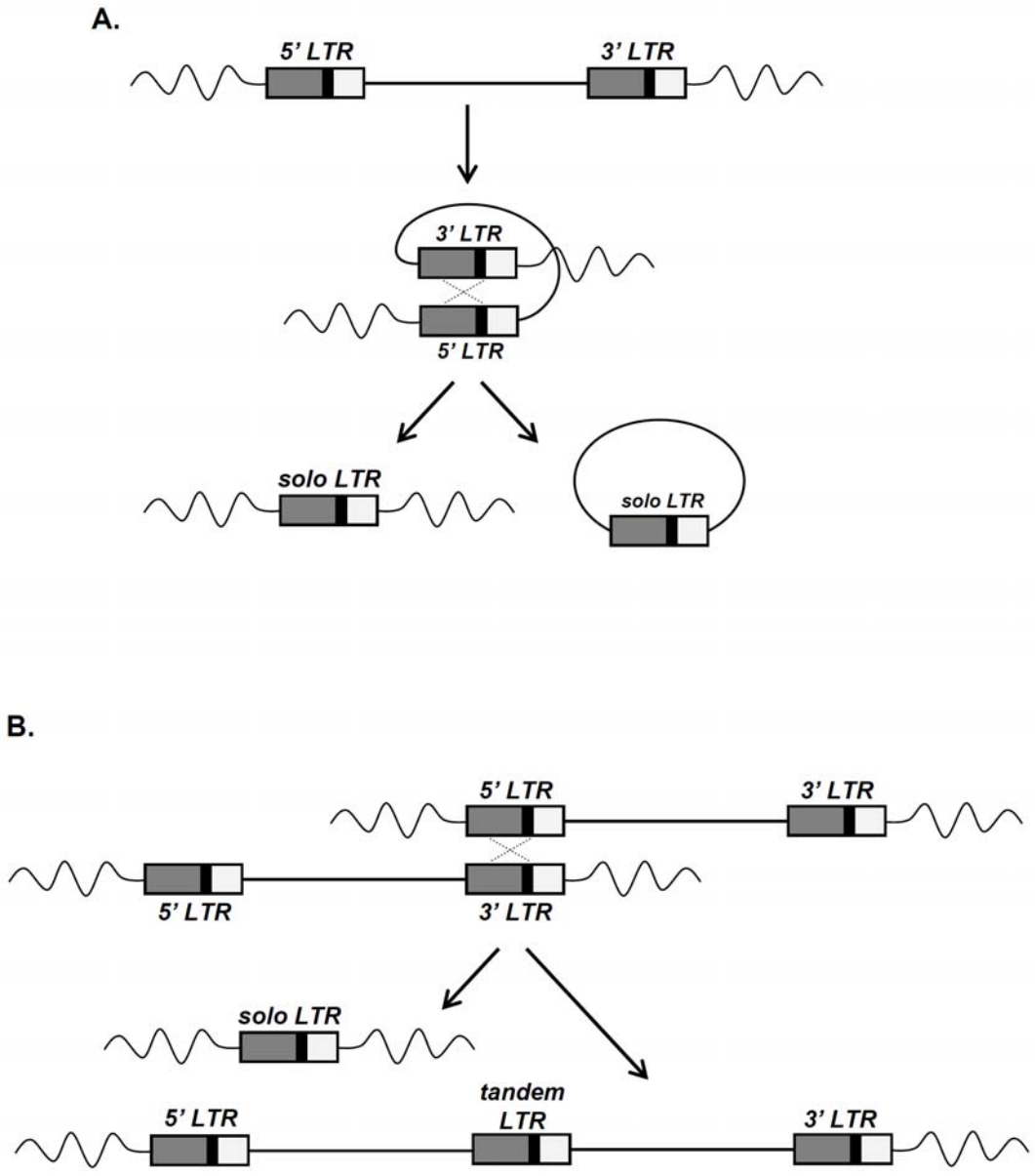


Figure 1.11. Events leading to formation of a solo LTR.
LTRs are color-coded to represent the U3 (grey), R (black), and U5 (white) regions; dotted lines between aligned 5' and 3' LTRs indicate the putative recombination site.
A. Formation of a solo LTR by recombination between paired 5' and 3' LTRs of a single provirus. The recombination generates a stable solo LTR in the host DNA, and a single LTR circle of the deleted internal region of the provirus.
B. Formation of a solo LTR from allelic recombination between sister chromatids, resulting in a solo LTR and a tandemly duplicated provirus at the homologous site.

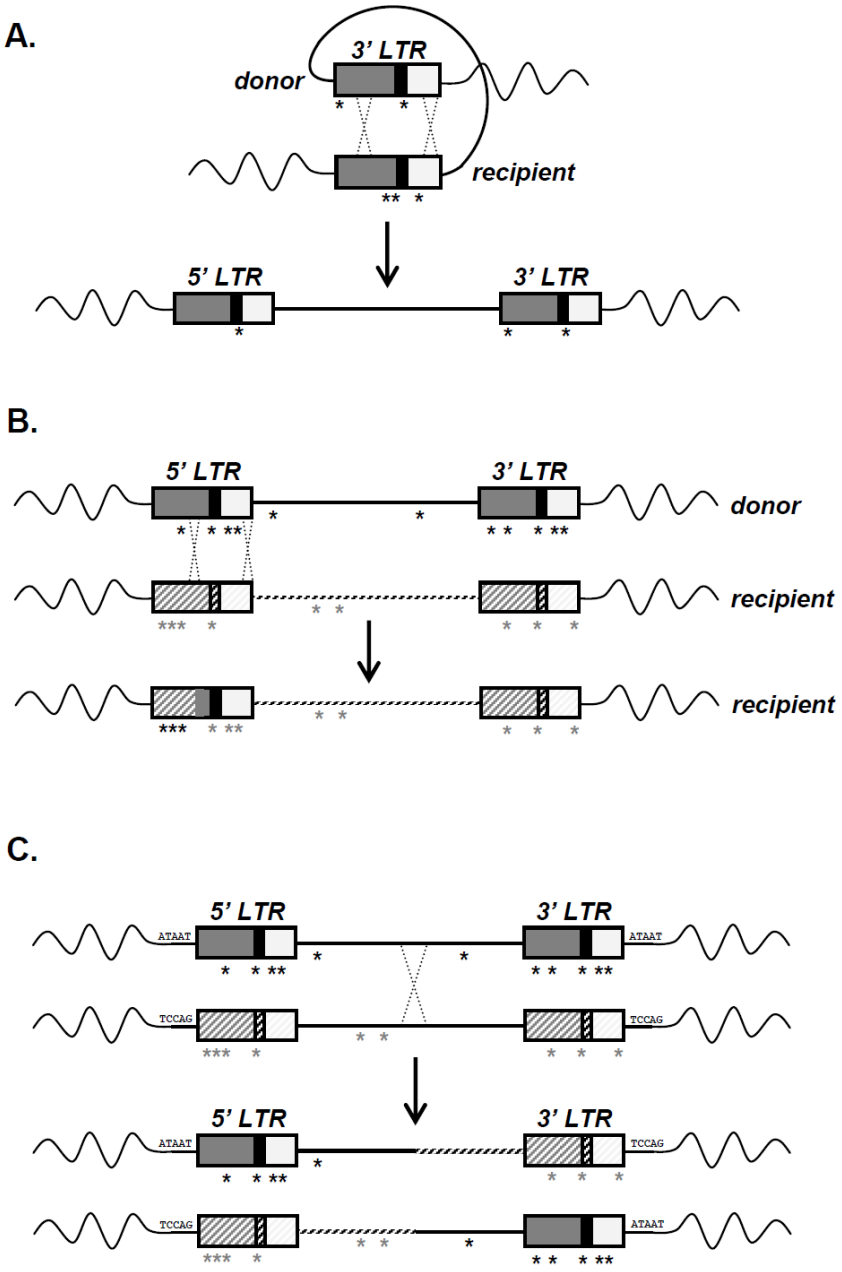


Figure 1.12. Effects of gene conversion or ectopic rearrangements between two proviruses.

The LTRs are detailed by the U3 (grey), R (black), and U5 (white) regions; dotted lines between elements indicate the recombination site. Asterisks indicate mutations specific to each element. Striping indicates distinct elements.

A. Gene conversion between the LTRs of a single provirus.

B. Gene conversion between two elements results in the uni-directional transfer of highly similar or homologous sequence, in which the information at the recipient site is completely replaced by that from the donor. The flanking and TSDs remain intact.

C. Ectopic recombination between two elements at distinct locations can lead to large-scale rearrangements. The flanking cellular DNA and TSDs are not left intact.

Of the total recognizable HERVs in the available genome database, up to ~90% are estimated to exist as solo LTRs [87, 88], the result of homologous recombination between the LTRs of a single provirus, or between LTRs at distinct loci [75]. Formation of a solo LTR by an intrachromatid recombination between paired 5' and 3' LTRs (Figure 1.11 A) is most likely to occur soon following the time of formation, when the LTRs are identical or highly similar in sequence. As the LTRs independently accumulate mutations over time, the rate of solo LTR formation drops dramatically, estimated at rates from 10-fold (in the case of a single nucleotide change) to >100-fold (2+ changes) [88, 89]. Solo LTR formation can also occur as a product of an interchromatid recombination between LTRs of allelic proviruses, potentially resulting in a tandemly duplicated provirus on one chromatid and a single LTR on the other (Figure 1.11 B) [76]. Probably the most well-characterized example of such an event is the polymorphic HML-2 provirus located at 7p22.1 (HERV-K108, or HML-2.HOM), which is found in tandem repeat, single provirus, or solo LTR forms at varying frequencies within the population [89, 90].

Aside from HERV integrations represented in the solo LTR form, many proviral elements exhibit evidence of DNA rearrangements having occurred without any loss of internal sequence, namely from gene conversion and ectopic recombination, also referred to as non-allelic recombination [63, 65, 91]. A gene conversion between proviruses involves the uni-directional transfer of the internal coding and/or LTR sequence information from a 'donor' element to a highly similar 'recipient' element,

resulting in the complete loss of the recipient's sequence information at that site (Figure 1.12). At least in mammals, the transfer usually involves short stretches (averaging about 1kb), however lengthier events, or successive transfer events can lead to the homogenization of larger regions [92]. Recombination leading to gene conversion may occur between allelic proviruses on sister chromatids or between elements in different chromosomal locations [75, 92]. Due to the nature of the conversion, the host flanking sequence generally remains intact, along with the 5-6 bp target site duplications, with the net outcome being the transfer of a 'copy' of a homologous sequence [63, 65].

Analysis of the HML-2 group has indicated at least two proviruses that have been subjected to gene conversion. The best-studied example has been from a provirus located at 20q11.2, in which the LTRs do not group as expected in a comparison with the homologous 5' and 3' LTRs in multiple species; the first ~70 bp from the 5' LTR are missing, presumably from deletion, and so could not be compared for sequence identity of the TSDs. What was most striking of this particular element came from the direct comparison of nucleotide differences between the 20q11.2 LTRs in the human with the homologous LTRs from other species, in which a clear tract ~300 bp in length is most similar to a distinct HML-2 element located on chromosome 9, highly suggestive of its gene conversion [65]. Another provirus with indication of gene conversion has been the K115 provirus (located at 8p23.1). The K115 LTRs do not cluster together [89, 93, 94], and the TSDs of this element are intact and identical, properties consistent with gene conversion. However the element has been integrated recently, as indicated from its detectable allele frequency in the population (averaging ~15% of sampled genomes [95], but detected as high as 43% depending on ethnicity of the samples tested [96]), without

the possibility of species comparison. Aside from HERVs, gene conversion and sequence homogenization is well-documented in several human cancers, and in repetitive genome regions such as segmental duplications [92].

Ectopic recombination between HERVs is distinguished from gene conversion in that it involves illegitimate, or non-allelic, recombination between closely related HERVs integrated at distinct chromosomal locations [97]. The analyses of such events has provided strong support that the participating elements were directly involved, and in fact, provided the template for the branchpoint of recombination. Depending on the location that the recombination is resolved, the outcome can have profound effects to the genomic structure of the host, resulting in insertions, deletions, or other large-scale chromosomal rearrangements. There is at least one well-studied example of a HERV-mediated illegitimate recombination that that is associated with a disease phenotype in humans. [98]. In this example, an intrachromosomal recombination between two elements of the HERV-I group on the Y chromosome results in a ~700 kb deletion of the region internal to the two HERV-I copies. The deletion, which is recurrent in humans, includes the gene for azoospermia factor A (*AZFa*, or *DYS11*). Males with this deletion are defective for sperm production; as a consequence this specific recombination is a common cause for male infertility [98].

Inter-element recombinations between HERVs have also been implicated in evolutionary changes in the host genome structure from the effective 'shuffling' of surrounding chromosomal regions [64, 65]. This type of recombination has been most documented, and appears to have occurred fairly frequently, among the HML-2 elements, with convincing evidence in nearly 20% of analyzed proviruses [64]. Similarly to gene

conversion, evidence of ectopic recombination can be inferred through the analysis of proviral 5' and 3' LTRs. At least 6 examples have been shown within the HML-2, in which these elements have unexpected clustering of cognate 5' and 3' LTRs, mismatch of the 5-6 bp flanking sequences, and discontinuity of flanking cellular regions (Figure 1.12) [64, 65]. As chromosomal rearrangements may have certain consequences to the host, evident from the association of recombinations between transposable elements with human diseases, they would likely be under a negative selection and removed from the population. This suggests inter-element recombinations may actually be fairly frequent, given that there are 6 examples observed within the HML-2 proviruses alone.

Regulation of cellular gene expression

While most, if not all, HERVs are heavily mutated and are no longer capable of generating functional particles, LTRs often have retained functional regulatory elements, such as cellular transcription factor binding sites and a poly-adenylation signal. Consequently, HERV LTRs are capable of affecting the transcriptional regulation of cellular genes through a variety of mechanisms [75, 77]. The intrinsic promoter activities of their LTRs can influence the transcriptional activities of nearby cellular genes, for example as a 'novel' promoter in its genomic context, of by acting to enhance the expression of a host gene in a tissue-specific or developmental-specific manner. Similarly, the LTRs can act to independently increased transcription from the native promoter of a host gene. HERVs have also been shown to be contribute to host transcripts as alternative promoters or termination sites, resulting in 'fusion' transcripts corresponding to *HERV-host* (or *host-HERV*) mRNAs. Similar transcripts may be generated by alternative splicing of host mRNAs to splice sites located within HERV gene regions, though this mechanism appears to be rarer in comparison to the former examples [77]. A few examples are described below (also diagrammed in Figure 1.13).

Promoter activities of HERV LTRs on cellular genes have been reported for several HERV groups [75, 77, 99]. One of the best-studied and most striking examples is the HERV-mediated tissue-specific regulation of amylase in the salivary glands of humans and other apes, but not in the Old World monkeys [100, 101]. The tissue-specific expression of this enzyme was shown to be mediated by a HERV-E integration within the 5' UTR of an ancestral copy of amylase that consequently activated for expression [100]. In another example involving a HERV-E element, a solo LTR acts

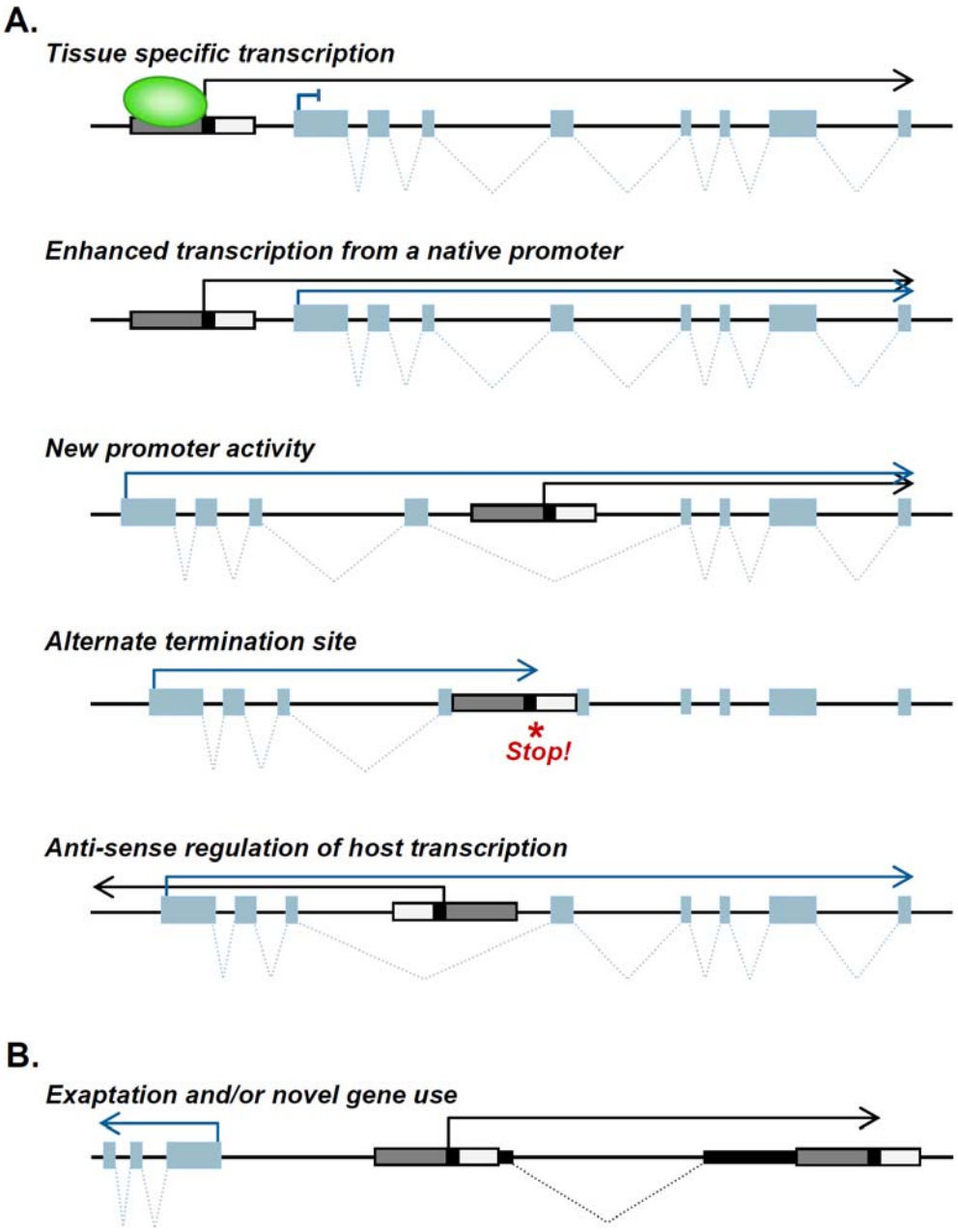


Figure 1.13. Potential mechanisms of HERV LTR-mediated transcriptional regulation within the host.
Depicted is a generalized LTR represented in various locales of a cellular gene. By virtue of the LTRs HERVs may modify the expression from such genes through mechanisms as indicated. Black arrow indicates transcription from the LTR, and the blue arrow transcription of the native gene. Blue boxes represent host exons, and dotted lines indicate splicing patterns within the LTR (A) or internal to a particular HERV (B). LTRs are labeled as in previous figures. *Stop!* indicates the premature termination of transcription of host transcripts by signals within the LTR.

to enhance the expression of Apolipoprotein C-1 (*APOC1*) from its native promoter [102]. The HERV-E LTR is situated about 300 bp upstream of the native *APOC1* promoter, and influences its activity *in cis*, resulting in a fusion transcript of the LTR and the *apoc1* mRNA. Thus, the LTR is not an enhancer itself but contributes to the transcription of the *APOC1* gene, and accounts for at least 15% of its expression in the human liver [102].

Another example of LTR-mediated expression is the tissue-specificity of leptin in humans. Leptin is usually expressed in adipose cells, however a HERV-K(HML-8) solo LTR situated upstream of the leptin gene contains several binding sites for transcription factors expressed in the placenta, and functions as an enhancer of leptin expression within placental tissues [103]. The case of leptin is made doubly interesting in that its receptor, *LEPR*, is also under the transcriptional influence of an LTR, although by a different mechanism. In this case, the integration of a HERV-K (HML-2) LTR within in the last intron of *LEPR* provides an alternative splice acceptor site, resulting in two splice variants of the *LEPR* gene, both of which are suggested to have functional physiological roles [104].

While these examples highlight normal physiological functions, other studies suggest involvement in disease. For example, a HERV-E integration between the 5' UTR of the human pleiotrophin gene (*PTN*) confers trophoblast-specific expression of a functional fusion transcript, owed in part to the presence of tissue-specific transcription binding sites [105]. HERV–*PTN* expression in trophoblasts is physiologically relevant, however pleiotrophin is also transforming in some cell types including, and is a key factor in the angiogenesis and metastasis of human melanomas [106]. HERV-*PTN*

transcripts are upregulated in tumors derived from human malignant trophoblasts (choriocarcinoma) but not in tumors from embryoblasts (teratocarcinoma), suggesting their possible involvement in the development and/or progression of choriocarcinoma [107].

Just as HERV LTRs can contribute to the transcriptional regulation of pre-existing cellular genes, one would predict the associated expression of proviral genes to follow suit. A compelling example of such expression is from the syncytins. The syncytin-1 and -2 genes have been respectively co-opted from the fusogenic *env* ORFs of distinct HERV-W and HERV-FRD proviruses, and have specific function in syncytiotrophoblast formation during placental development [108, 109]. Such *env*-derived fusogenic gene 'pairs' have since been identified in other placental mammals, in each case having been independently appropriated by the host (for example, see [110], and also reviewed in [111]). The syncytins have also been implicated in disease. Syncytin-1 has been detected in patient tissues and cell lines derived from breast tumors. In these tissues, syncytin-1 has been shown to be capable of mediating fusion between cancerous and adjacent normal tissues [112], and could possibly be involved in promoting tumorigenesis in some cancer types, though this remains to be clarified.

Regulatory effects of human-specific LTRs

Alternate promoter use may lend a certain specificity in either the tissues involved, or in the timing of the expression patterns of nearby cellular genes. Because inducing a specificity to the expression allows for specific changes in expression, regulation by alternate promoters has been hypothesized to influence species diversity

and evolution [113, 114]. Along these lines, the HML-2 group of elements has been given special attention as the only HERV group with human-specific loci. Individual human-specific LTRs (including solo LTRs) from the HML-2 have been investigated for promoter activity under the hypothesis that they may have contributed to unique gene expression patterns in humans following the divergence from the chimpanzee [105, 107].

In a primary study including human-specific LTRs from both proviral and solitary loci, transcriptional activity was analyzed in tissues from human testis, and promoter activity was observed for at least 78 human-specific elements, or about half of those tested [80, 85, 115, 116]. Subsequently, the authors compared the activities of these LTRs in tissues corresponding to 'healthy' testis (testicular parenchyma), and to the corresponding testicular cancer (seminoma). For those LTRs in gene-rich regions (either internal to the gene or within ~10 kb), several were found to be up- or down regulated with specificity toward either tissue type. The most striking examples included a ~10-fold increase in seminomas for the promoter activities for LTRs proximal to genes that encoded the *AND-1* DNA binding protein and the transcriptional enhancer *CEBPZ*, and there was ~5-fold increase in transcription from an LTR proximal to *SLC48*, a solute carrier protein. The opposite situation was also observed, for example an LTR near *LIPH1*, encoding the a membrane-bound lipase precursor, was observed to have a ~5-fold increase in normal testicular tissues [85].

Of the total LTRs reported as 'within-genes', the vast majority were located within introns, and in the anti-sense orientation, suggesting their possible affects to their respective host genes through an anti-sense regulatory mechanism (for example see Figure 1.13 A) [85]. Analysis of the transcriptional levels of two cellular genes, *IFT172*

(an intraflagellar transport protein) and *SLC48* (also above), in either testis cell type, confirmed that the increased activity of the respective intronic LTR was correlated with a decrease in the transcript levels of either cellular gene [85]. How each of the remaining cellular genes responds to the nearby change in promoter activities of these human-specific LTRs remains to be seen.

These examples highlight the how the differential promoter activities of some human-specific HERVs can affect the expression of nearby host genes in a cell-type specific manner. A human-specific change in transcriptional regulation has the potential for novelty, as the change represents an new evolutionary acquisition that is lacking in chimpanzees. However, also implied is the potential for changes in cellular gene expression that may have negative consequences to the host species. As the mechanisms and consequences of transcriptional regulation of human-specific elements are further characterized, we will likely gain more knowledge of how these elements may contribute to differences in patterns of expression between tissue types.

The HML-2 group of proviruses

HML-2 is one of ten phylogenetically distinct sub-groups (HML-1 through HML-10) belonging to the larger 'umbrella-like' group of HERV-K [32, 38]. The HML-2 were first described in the 1980's as the 'human MMTV-like' elements that had been identified through low stringency hybridizations of MMTV-specific probes to human genomic DNA [117]. The MMTV-like sequences, sometimes referred to as 'HMTV', or 'human MMTV', had ~95% identity to MMTV, and few inactivating mutations. Following the initial reports of MMTV-related sequences in humans [117], the first full-length provirus, HERV-K10, was cloned and characterized as a human endogenous retrovirus by Ono *et al*. in 1986 [118, 119]. The potential for a human cancer-causing agent analogous to MMTV in mice led to much of the initial searching and characterization of the HML-2 group.

Members of HML-2 phylogenetically group with the *Betaretroviruses* (formerly type-B and -D) based on the most conserved region of the *pol* gene, and are thus referred to as '*Beta*-like' [32]. Other retroviruses in this genera include mason Pfizer monkey virus (MPMV), Jaagsiekte sheep retrovirus (JSRV), and MMTV –the latter two viruses having both endogenous and exogenous elements [37]. Within the HML-2 group, the env ORF has been shown to be highly conserved, providing support that the primary mode of germline replication of the HML-2 has been through germline infection [70, 93].

The identification and characterization of HML-2 (as has been the case for other endogenous retroviruses) has been ongoing. This is reflected in nearly all successive reports since their discovery, given the continued advancement in technologies and present knowledge in the field. As a result there has been a considerable

range in the estimations of the total numbers of individual HERV groups, not excluding HML-2 as well. The most recently reported HML-2 representatives have totaled around 35 in 2003 [94]; 54 were reported in 2006 [120]. In the Results and Analysis in Chapter 3 is presented an updated collection of HML-2, and given those results, we report ~89 proviruses and ~944 solo LTRs from the HML-2 (R.P.S., J.H.W., *et al*, in preparation).

HML-2 genome organization and group-specific features

The genomic structure characteristic of the HML-2 group is depicted in Figure 1.14. HML-2 proviruses are ~9.4 kb in total length, with flanking LTRs of ~970 bp. In addition to the canonical *gag*, *pro*, *pol*, and *env* genes, these elements have been shown to contain a dUTPase domain within the *gag* ORF. Finally, HML-2 proviruses encode one of two additional proteins, namely, Rec or Np9, whose transcripts are generated through the subgenomic splicing of *env* mRNAs [121, 122]. An intact provirus has been shown to draw attention to the viral genes and the specific deletion of 292 bp (Δ292 bp) that spans the *pol*-*env* boundary. The deletion, shaded in Figure 1.14, is present in a subset of elements within HML-2, and has functional consequence for those proviruses. As a result, the Δ292 bp has been used as a basis to structurally classify each provirus into one of two subtypes based on the presence (type 1) or absence (type 2) of the deletion [122].

The type 2 HML-2 proviruses encode the accessory protein Rec. The Rec protein, previously referred to as 'cORF', has been shown to have functions analogous as the HIV Rev and HTLV Rex proteins, in that it associates with unspliced and partially spliced viral mRNAs and promote their transport from the nucleus to the cytoplasm [123]. In HML-2, Rec does this through association with the Rec Responsive Element

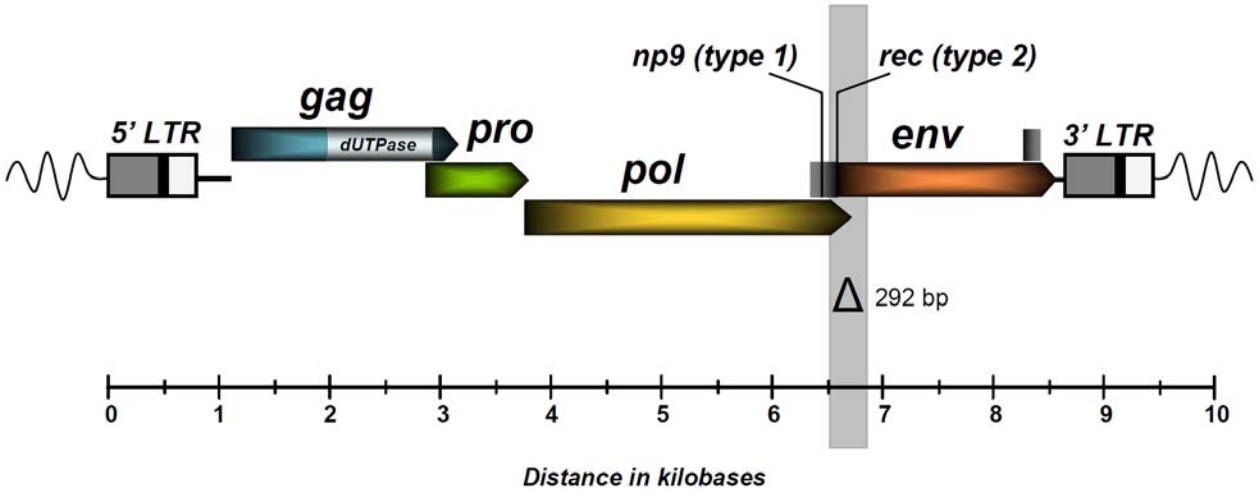


Figure 1.14. Genome Structure and characteristic group-specific features of the HML-2 proviruses.
The typical genome structure is shown for an HML-2 provirus, shown to scale. Positioning of the ORFs in the internal portion of the provirus correspond to reading frames. The dUTPase encoded at the 3' end of gag is labeled and shown boxed. Gray shading at the pol-env boundary indicates the coordinates of the 292 bp deletion used to differentiate the HML-2 according to subtype: type 1 carry the deletion and type 2 do not. Type 1 elements encode the *np9* reading frame, and type 2 encode the *rec* gene. The splice sites corresponding to the Np9 or Rec proteins are delineated and labeled. The *rec* ORF is shown in gray with the other canonical genes. LTRs are as in previous figures.

(RcRE) structurally encoded within the LTR (also indicated in Figure 1.15) [121, 124]. Transcripts for Rec are generated through subgenomic splicing of *env* mRNAs at minor splice sites located within *env* [121]. The *rec* alternative splice site is deleted within the Δ292 bp of type 1 HML-2 proviruses, resulting in transcripts for a ~9 kDa fusion protein referred to as Np9 [121, 124]. Both Rec and Np9 are detected in human tissues and each is associated with abnormal or tumorigenic cellular conditions, in particular in breast tumor tissues [122, 125, 126]. Implications for either protein with disease are discussed further below.

Irrelevant to subtype, HML-2 proviruses have been shown to encode functional dUTPase motifs within Gag [127]. This motif is present in several retroviruses, including the *Betaretroviruses* MMTV, MPMV, and JSRV, as well as non-primate *Lentiviruses* and other HERV groups [127]. The function of dUTPase is to hydrolyze dUTP to dUMP, resulting in the decreased chance of misincorporation of dUTP into newly synthesized DNA. In the case of retroviruses, the presence of a dUTPase is thought to confer a benefit to a replicating virus by reducing mutagenic effects of dUTP incorporation during reverse transcription [127].

The canonical HML-2 LTRs are 968 bp in length and contain several features that are characteristic across the group (structurally diagrammed and labeled in Figure 1.15, *upper*) [118]. These include a putative TATA-containing promoter, polyadenylation signal and site for transcription termination. Also characteristic of HML-2 LTRs is the location of a ~75 bp enhancer region near the 5' end that contains binding sites for multiple host transcription factors, including the YY1 binding site shared among all HML-2 [128]. Just upstream of the enhancer core is part of a glucocorticoid-responsive

element (GRE) that, in combination with the enhancer core, contributes to the formation of a putative binding site for progesterone receptor complex [129]. Binding sites for several host transcription factors and protein complexes are also contained within this region, and have been shown to enhance HML-2 expression in cells lines derived from germ cell tumors [128].

Most of the HML-2 that have intact LTRs share the canonical promoter motifs described above, and although the HML-2 transcription has been referred to with reference to these sites, two recent reports provide evidence for at least some HML-2 to have 'TATA-less' promoters (Figure 1.15, *lower*) [116, 130]. In particular, the 2011 study by Fuchs, *et al*. mapped the transcription start and stop sites for the K108 (located at 7p22.1) and K18 (at 1q23.3) elements. Aligning multiple transcript reads generated by 5' and 3' RACE indicated a major start and stop site, placing the R region at ~800-900 bp from the LTR start, and making it unlikely those transcripts arose from TATA-dependent activity [130]. In support, abrogating the TATAAT sequence itself had no decrease on the transcriptional activity from the LTR [130]. Also, multiple binding sites were identified for transcription factors, such as SP1/SP3, that have roles in other TATA-less promoters (Figure 1.15, *lower*). The presence of each binding motif was shown to correlate to the transcription of a dozen known HML-2 LTRs, providing support for the requirement of certain host factors for transcription [130].

The results of these analyses suggest that HML-2 expression in some human tissues occurs in the absence of the putative TATA promoter, and is in agreement with several reports of other HERV groups. As the experiments were performed in cell lines derived from germ cells and pluripotent embryonic stem cells, the promoters could be subject to

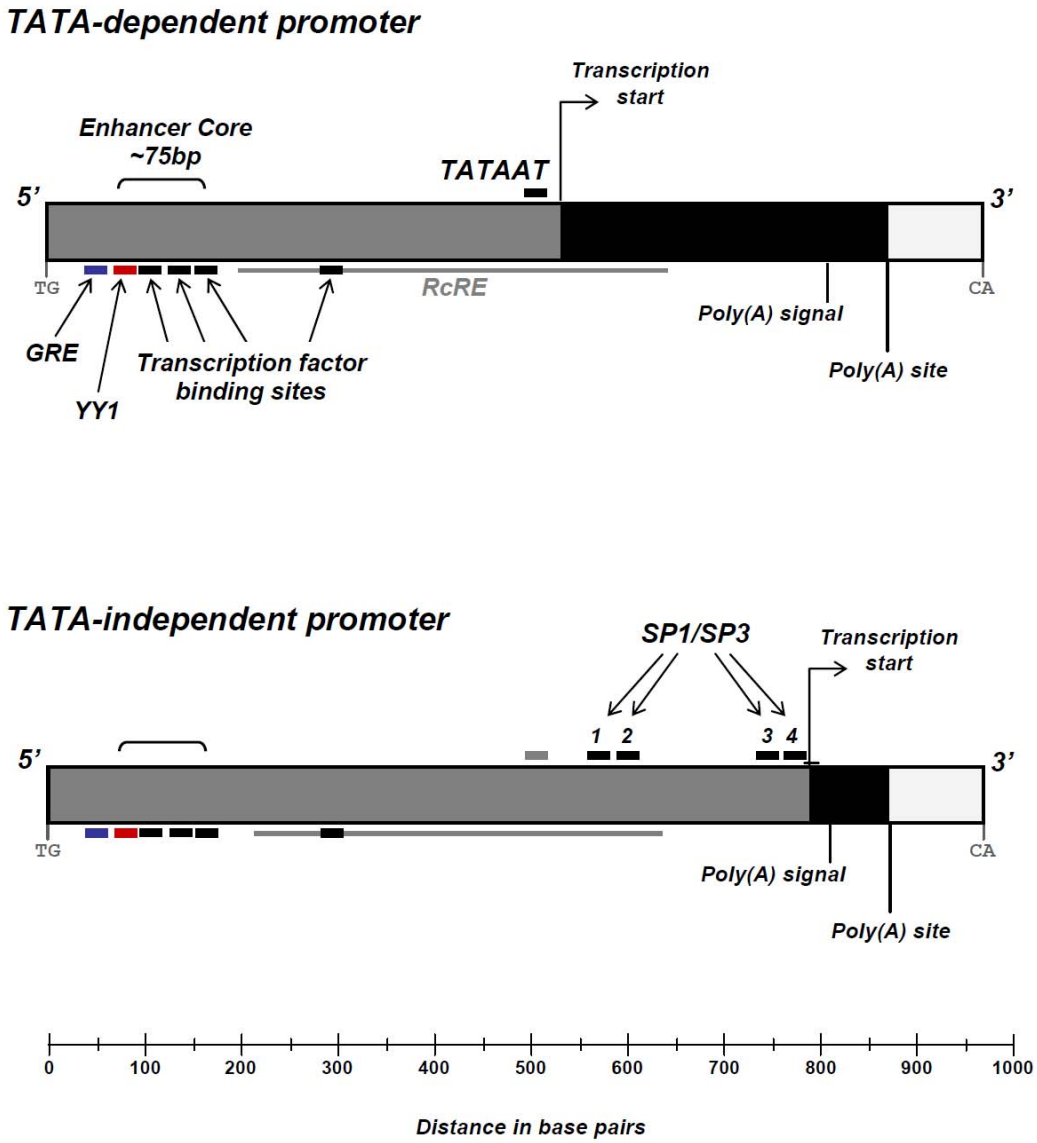


Figure 1.15. Representative promoter types within the HML-2 group.
Upper: A TATA-dependent promoter is shown to scale with the locations of transcription signals and transcription factor binding sites. The position of the TATA box is indicated with black line and labeled. The GRE and YY1 transcription factor binding sites are labeled and shown in blue and red lines, respectively, flanking the edge of the enhancer core. Labeling is as follows: the TG/CA represent the flanking canonical dinucleotides; RcRE indicates the Rec-responsive element; transcription start for each promoter is arrowed; relative position of the poly(A) signal and site are labeled.
Lower: A TATA-less promoter is shown. The binding sites for the TATA-independent transcription factors SP1 and SP3 are indicated. All labeling is the same as for the TATA-dependent promoter.

different transcriptional regulation. As the TATA sites for most HML-2 elements are present and intact, and the authors suggest the ability to infect and replicate in different cell types requires a switch from a TATA-dependent to a TATA-independent promoter. Support for this hypothesis is offered from the 2006 study by Kovalskaya, *et al*. [116], in which two species of HML-2 transcripts were confirmed from human testicular parenchyma. In this scenario, HML-2 may have retained at least two promoter activities, and further analysis may offer some explanation to the increase in transcriptional activity of the HML-2 in distinct cell types.

HML-2 activity in humans

Among HERVs, HML-2 is the most conserved in sequence and the only known HERV group with activity following the *Homo-Pan* divergence. The HML-2 group is represented in total by ~90 full-length or near full-length elements, of which ~25 are human-specific proviruses. Of the species-specific proviruses, ~11 are polymorphic, as indicated from the detection of at least the provirus and solo LTR alleles at the site. About 5 have been shown to be insertionally polymorphic, as those particular integration sites are not yet fixed among humans, and the empty, or pre-integration site, is still detected [89, 90, 95, 131].

Although all described HML-2 proviruses have apparent inactivating mutations, most of the human-specific proviruses contain at least one intact ORF, and at least two integrations, namely at 7p22.1 (K108) and 19p12 (K113) have intact reading frames in their full-lengths [90, 95]. The K-108 Env is capable of mediating infection in pseudotype with VSV-G, and is one of just 16 HERV-encoded envelope proteins with described fusogenic activity [132]. HERV-K115 encodes an enzymatically functional reverse transcriptase [132], and expression of Gag from HERV-6q14.2 (K109) or 22q11.21 (K101) is sufficient for particle formation [133]. Particles released from Tera-1 cells were recently demonstrated to efficiently and selectively package transcripts from the K101 provirus, providing evidence that at least some proviruses have retained active packaging signals [134].

Of the described HML-2 proviruses, K113 appears to be the most recently known integration, as the comparison of its paired LTRs gives an integration time within the last ~100,000 to 200,000 years [95]. Furthermore, this particular provirus was

recently shown to have retained the capacity for the formation of virus like particles having morphological features consistent with other *Betaretroviruses*, though none were capable of infection [135, 136]. Noteworthy in this regard, no 'naturally occurring' HERV of any kind has been observed with infectious capacity. However, two engineered proviruses generated with reference to the consensus nucleotide sequence of the most conserved HML-2, have been shown to be capable of replication and *de novo* integration [133, 137].

Given these collective observations, the HML-2 group is regarded as the most likely of HERVs with the potential for activity. Sequence comparisons indicate HML-2 genes, including the *env* gene, have been under purifying selection, suggesting entry into the host genome via independent infection events [93]. Also, the HML-2 rate of germline insertion is suggested to have been constant since the human divergence from chimpanzee based on standard population genetic models [88]. Given the conservation among the most recent HML-2 proviruses, and their formation within relatively recent evolutionary time, it has been suggested that a pool of replication-competent HML-2 elements exists within our population.

Patterns of expression and implications for human health

Retroviruses were first discovered as tumor-inducing agents in animals and have historically been associated with cancer. In animals a wide variety of cancers have been attributed directly to retroviral infection from horizontally transmitted viruses, however there are also well-studied cases in which ERVs with replicating exogenous variants are directly involved in tumorigenesis [29, 40, 138]. Classic examples are from

the endogenous MLV and MMTV, which cause leukemia and breast cancer in mice, respectively, both by oncogene activation following the formation of novel proviral insertions [12, 138]. In the case of MLV, infectious viruses capable of *de novo* integration are formed as recombinants from at least three endogenous MLVs [12]. Endogenous MMTV is upregulated in the mammary glands of lactating mice, which may likewise lead to the formation of infectious viruses and subsequent *de novo* proviral formation [12]. More recently have been the identifications of additional murine ERVs able to generate clonal proviruses in tumor cells for example the so-called melanoma-associated retrovirus (MelARV) [139, 140]. Aside from murine examples of cancer-causing ERVs, another example has been from JSRV, which causes a contagious lung cancer in sheep through oncogenic affects of the encoded Env protein [141]. *In vivo* expression of the JSRV Env is sufficient to induce lung cancers in sheep and mice [142].

Identification of ERVs in animals that have pathogenic or tumorigenic effects has undoubtedly motivated the search for HERVs with similar properties in humans, although such elements have yet to be described. Numerous reports have described the presence of transcripts, proteins, and particles from HERVs in certain cell-types and tissues that correspond to a variety of human diseases. This has particularly been the case for proviruses belonging to the *Beta*-like HML-2 group, although it is worth noting that transcripts are observed for other HERV groups including, but not limited to, some members of the Class I HERV-E and HERV-H groups [143]. Expression of HML-2 proviruses has been well-documented to be up-regulated in tissues associated with several diseases, one of the most common being from breast cancers [129, 144-148],

germ cell tumors [134, 149-151], melanomas [152-154], neurodegenerative disorders [24, 25, 27], as well as during HIV infection [155-157].

The expression patterns of individual HML-2 have been shown to have differential activities, such that distinct HML-2 proviruses are more strongly upregulated in some cell types but not others [151, 158]. For example, the most frequently detected transcripts in germ cell tumors have been from 22q11.21 (K101), 11q23.3, and 1q22 (K102), in that order, however 19 others are detected at much lower frequencies in these tissues [151]. Likewise, in brain tumors, transcripts from the 5q33.3 provirus (K10) account for >70% of detected cDNAs, though transcripts from at least 9 others are also detected. These observations indicate a more 'global' up-regulation of the HML-2 proviruses, a possible explanation for which is following the deregulated expression of transcription factors present within different tumor tissues [151]. Of note, all but 6 human-specific proviruses (and accounting for all described polymorphic integrations) have been detected in such analyses, with the indication that other uncharacterized polymorphic integrations within humans would likely be expressed in such tissues.

In the Results and Analyses Part II of this document (Chapter 4), we present a study of polymorphic HML-2 in human genomic DNAs, in the context of two diseases with demonstrated up-regulation of HML-2, breast cancer and schizophrenia, and evidence for either case is detailed in the following sections.

HML-2 expression in breast cancer

The HML-2 group has been connected to breast cancer from observations of increased transcriptional activities of HML-2 proviruses both from breast tumor patient

biopsies and in cell lines derived from breast tumor tissues [122, 129, 144, 146-148, 159-161]. The expression of HML-2 transcripts and proteins have been shown to be specifically up-regulated in the mammary tumor tissues but not adjacent epithelia of most tested breast cancer patients [122, 146, 147]. For example, matched tissue analyses have identified full-length (i.e. *gag*) and spliced (i.e. *env* and its spliced variants) transcripts specifically expressed within primary mammary tumor samples, from which cDNAs shared >97% identity to the 1q22 (K102) provirus [146, 147]. The 2008 study by Flockerzi *et al*. provided support from the identification of >70% of breast tumor transcripts having originated from the same provirus [151]. Also have been the reports of RT activity, and the detection of Env and Gag using antibodies specific to the HML-2 expressed proteins [148, 162], and the release of RVLPs from the human breast tumor-derived cell line T47D [145]. The release of RVLPs is also characteristic of other types of cancers, for example in teratocarcinomas [134].

Both type 1 and type 2 HML-2 proviruses are expressed within breast tumor tissues, as demonstrated from the detection of spliced transcripts originating from either subtype [144, 146, 147, 151]. As mentioned, most HML-2 type 1 proviruses encode a reading frame for the Np9 protein due to a deletion within *env*, whereas type 2 proviruses encode the reading frame for the Rec accessory protein; the expressed proteins share the first 14 N-terminal amino acids (also refer to Figure 1.14). Although *rec* transcripts are found in normal and cancer tissues, *np9* mRNA has been shown to have specific expression in tumor tissue, as was first observed in tissues from mammary carcinoma biopsies [125]. Both proteins were subsequently shown to interact with the promyelocytic zinc finger protein (PLZF), a transcriptional repressor of cellular genes that include the c-

myc proto-oncogene, and has suggested function as a guardian of stem cell pluripotency [126]. Furthermore, Rec-expressing cells show increased proliferation, decreased apoptosis, and can induce tumor formation when injected into mice [163]. Np9 alone has been shown to bind and functionally interfere with the ligand of Numb protein X (LNX), a ubiquitin ligase of the Numb/Notch pathway involved in Ras signaling [125]. As such, both Rec and Np9 have been implicated as oncoproteins however whether they may be capable of causing cancer is still unclear.

HML-2 expression in neurodegenerative disorders

HERV expression has also been associated with neurodegenerative disorders, in particular schizophrenia and bipolar disorder [24-27, 164, 165]. There are a growing number of reports of HERV expression in the form of transcripts, proteins, and particles in patients diagnosed with schizophrenia. The most studied HERV with respect to neurodegenerative disorders has been from the HERV-W group. HERV-W transcripts have been observed in brain tissue specimens and in plasma of patients diagnosed with recent-onset schizophrenia [24-27]. Particle-associated RNAs originating predominantly from HERV-W loci, but also having high sequence similarity to HERV-K and other HERV groups, have also been isolated from the cerebrospinal fluids and brain tissues of such patients [24, 27]. Significant levels HERV-W transcripts have also been detected in peripheral blood mononuclear cells of patients with recent-onset diagnoses (100% tested) as compared to chronically affected schizophrenics (~26% tested) [25].

The HML-2 group has also been associated with schizophrenia. A previous study by Frank et al. in 2005 reported differences in the transcriptional activities of a

subset of HML-2 in a *pol*-based microarray analysis of cDNAs from post-mortem prefrontal cortex tissues of schizophrenics and matched controls [25]. Tested HML-2 loci included K10, K108, 10p14, and an older HML-2 provirus located within the 8p23.1 chromosomal band and proximal to K115. Transcripts from K10 were found to be significantly over-represented in schizophrenia ($p_{K10} < 0.025$) and bipolar cases ($p_{K10} < 0.007$) compared to matched controls, suggesting differential expression from a subset of HML-2 loci [25]. Other *pol*-based studies using primers directed to the general HML-2 group have shown upregulation of HERV-K transcripts within brain tissues of schizophrenic individuals [27].

Only recently have HERV transcripts been analyzed individually to characterize the transcriptional activity from specific proviral loci within tissues from diseased and healthy tissues [166]. In a 2008 study from Flockerzi, *et al*., at least 16 individual loci were found to be transcribed within various brain tissues including the prefrontal cortex [151]. From the total representative cloned cDNAs, the HML-2 proviruses located on 3q21.1 (HERV-K(I)) and 5q33.3 (K10) were detected with the greatest frequency, making up ~50% of total clones [151]. Interestingly, clones from 22q11.21 (K101) were detected quite rarely (1 of 654 total) in brain tissues but have been shown to be strongly upregulated in tissues derived from germ cell tumors, suggesting this particular provirus is differentially expressed in different tissue types. As with HML-2 expression in breast cancer, proviruses of both subtype have been shown to be up-regulated in brain tissues of schizophrenics; the expression patterns of *rec* and/or *np9* mRNAs, or the encoded proteins, is not yet clear.

Potential for HERV involvement in human disease

Whether expression of HERVs has pathogenic consequences in humans remains in question. HERV transcripts have been detected in every tissue type analyzed thus far [167], and although their expression in several diseased tissues differs from healthy controls, HERV RNAs appear to be a part of the general transcriptome. The question remains: Do HERVs have any significance to human health? As has been documented in animals, at least two mechanisms exist from which HERVs may cause disease: 1. By direct oncogenic effects of HERV-encoded proteins; 2. From proto-oncogene activation following *de novo* integrations in somatic cells, either as a result of the formation of a recombinant virus capable of infection, or from the expression of a single replication competent provirus. Both mechanisms have been documented in animal models, as was explained in the cases for JSRV in sheep (oncogenic properties of *env*) [151], and endogenous MLV and MMTV in mice (insertional mutagenesis) [168]. Alternatively, upregulation of HERV expression could be the result of an altered transcriptional environment of the diseased host cell, in which case individual HERV loci may provide useful markers for the disease.

There is some potential from the HML-2 to encode oncogenic proteins, as was described above for the Rec and Np9 accessory proteins expressed by either subtype. Perhaps an attractive premise of such an explanation of association with disease is that no mobility (i.e., *de novo* integration) is required. However, the HML-2 are also considered to be the most likely candidates for the existence of an infectious provirus, or 'pool' of proviruses, that may have retained infectious capacity. Because a replication competent provirus would likely have negative, or deleterious, effects to the host, it is expected the

element would be detected at low frequencies within the population. In theory, the infection and *de novo* integration of such an element could cause disease in a fraction of individuals carrying the provirus as in animal models of insertional mutagenesis. Conceivably, infectious viruses could be encoded within a single intact provirus, or generated through recombination of multiple proviruses. Notably, formation of such a retrovirus has been demonstrated from the *in vitro* 'recombination' of three HML-2 loci having the Gag of K109, the K108 Env, and the K115 Pol [133]. However, a replication-competent 'naturally occurring' provirus has yet to be observed from any HERV group. Nevertheless, the absence of observed elements capable of infection does not exclude the possibility of their existence.

CHAPTER 2

Materials and Methods

***In silico* data mining for proviruses belonging to HML-2**

To identify the chromosomal coordinates of HML-2 proviruses in human DNA, we searched the most recent genome build (GRCh37/hg19, February 2009) using the UCSC BLAT program [169] for sequences related to the full-length nucleotide sequence of the K113 provirus (AY037928) [131, 169]. Subsequent searches were performed with the engineered consensus sequences HERV-K_{CON} [137] and K-PHOENIX [133] to confirm returns. The DNA flanking each individual 'hit' was manually searched for sequence with high similarity to prototypical HML-2 sequences as determined by the RepeatMasker program in the UCSC genome browser [170]. For each identified locus, complete nucleotide sequences were generated by extracting and concatenating the internal and LTR proviral segments. Additional BLAT searches with individual K113 genes (*gag*, *pro*, *pol*, and *env*) were performed to further identify HML-2 elements within the available genome. Again, the consensus HML-2 sequences were BLAT searched in the same manner for support in our strategy and findings. Complete sequence reconstruction was performed as for the full-length elements, with the minimum criterion for a provirus being the presence of an LTR and a "hit" matching >50% of the length of a full gene, or two proximal genes with >50% frames and no LTR. All full-length sequences were initially aligned to K113 using ClustalW [171], and manually edited in BioEdit v.7.0.9.0 [172].

The full-length sequences for the HML-2 proviruses located at 10p12.1 (K103; published as a solo LTR) and 19p12 (K113; not included in the current or any previous genome builds) were from NCBI (accession numbers AY037928 and AF164611, respectively). The 12q13.2 provirus has also been included in the published

human genome builds as a solo LTR; the locus was shown to have polymorphic alleles in previous work by Belshaw, *et al*., in 2005. We sequenced the full-length 12q13.2 provirus for these analyses, in methods described below (see following section 'Amplification and sequencing of the 12q13.2 provirus').

The K105 provirus was identified on an unaligned chromosomal region, originally said to be identified within centromeric sequence [173]. The provirus was said to be a type 1 element (containing the Δ292 bp diagnostic of the subtype), and detected in humans and chimpanzee but not gorilla [173]. Despite a number of comprehensive searches, we were unable to locate the full-length sequence in any database, and the sequence was obtained circuitously. We first verified the location and flanking sequence of the human K105 by BLAT of the 5' and 3' LTRs (AH008413.1) available through NCBI, and extracted the K105 solo LTR with ~1kb flanking sequence in either direction. Using this sequence, we mined the chimpanzee database and identified the corresponding provirus within a chimpanzee BAC (at position 74813; AC195095.2). Subsequent BLAST search of this provirus and its flanks returned a recently deposited HML-2-like provirus 'K111' (GU476554) with 99% identity to the chimpanzee K105 query. Subsequent comparison of the human K105 5' and 3' LTRs revealed a similar level of sequence identity (~99%), in support that the K111 is in fact the K105 provirus. In further support, we observed the same number of base changes between LTRs (38 in total) of the K111 element and from the previously reported K105. Thus, we have included the full-length sequence within these analyses as the K105 human provirus.

In the published human genome builds, the pre-integration site is present for the K113 site at 19p12. Sequences of ~150 bp in either direction of the K113 integration

site were previously cloned [95] and deposited in the NCBI database. A simultaneous BLAT against the human genome was performed with the 5' and 3' flanking sequences to confirm the K113 integration site and to infer the orientation of the provirus for subsequent analyses (also refer to Table 3.1).

Amplification and sequencing of the 12q13.2 provirus

The 12q13.2 site was identified in a 2005 study by Belshaw, *et al*. [131]. For about 100 human-specific HML-2 sites that were published as solo LTRs, the authors designed flanking primers and screened human genomic DNAs for the presence of either remaining allele. Pre-integration sites were confirmed for 8 solo LTRs, and a single site was characterized to further carry a provirus in some individuals: 12q13.2. We confirmed the location of the solo LTR by simultaneous BLAT searches of the published 12q13.2 site-specific flanking genome region [131], and the K113 full-length sequence. Primers were designed complimentary to the genome regions flanking of either direction within 1kb of the 12q13.2 site using the Primer3 v.0.4.0 interface [174]. Primers are provided in Table 2.1.

Using the 12q13.2 locus-specific primers, a sample of ~20 human genomic DNAs [175] were screened in two PCR reactions to detect all three alleles. The first reaction was with 12q13.2-flanking primers to detect the presence of either the solo LTR or empty site alleles. The second reaction paired the 12q13.2F primer with an HML-2-specific primer (HML-2R: 5′-CTCGAGCGTACCTTCACCCTAG-3′) to detect the 12q13.2 5'LTR, indicating the presence of the full-length provirus. The HML-2 specific primer was complementary to a highly conserved and specific ~35 bp region downstream of the

5' LTR edge (bases 1017-1039 of K113), and was designed using an alignment of the full-length HML-2 with highest percent identity to K113. PCR reactions were analyzed by gel electrophoresis, and a representative product from each reaction was excised and sequenced in both directions to confirm the specificity of the amplification.

A sample with two proviral copies at 12q13.2 was selected for sequencing to reduce background amplification of either the solo LTR of empty site products. The provirus was amplified in 4 overlapping segments using conserved primers internal to the provirus [173] paired with either 12q13.2F or 12q13.2R (PicoMaxx, Stratagene). PCR products were purified (Qiagen) and sequenced to ~6x coverage using a described HML-2 primer set from Barbulescu, *et al.*, from their the 1999 analysis [173] of the human-specific HML-2 (provided in Table 2.2). Individual sequence traces were manually edited and aligned to the K113 nt sequence in BioEdit v.7.0.9.0 [172], and the consensus sequence introduced into the full-length HML-2 alignment.

Table 2.1. Primer strategy for amplification of overlapping 12q13.2 proviral products.

	Forward (5' → 3')	**Reverse (5' → 3')**	**Exp. size (bp)[a]**
Primer			
12q13.2F:	CGGAGAATTCCACCTTCAAA		na
12q13.2R:	TGCATTGTGGTCATCCATTT		na
Fragment			
5′LTR	12q13.2 F	TTGTTCTGGAAACCATGGGC	1377
5′LTR-*gag*	12q13.2 F	GAGCGGGCATGGTCATTTCC	3880
5′LTR-*pol*	12q13.2 F	AAGACCAATCTGCCATGCAC	5258
pol-3′LTR	AGAGGTTGCCAATGCTGGAC	12q13.2 R	5343
env-3′LTR	CCACTCCTCAGATGCAACTT	12q13.2 R	3678
3′LTR	AGGAGTTGCTGATGGCCTCG	12q13.2 R	1586

[a] Size was estimated by adding the total # bp from the internal primer to the appropriate LTR edge, and the paired flanking primer to the same edge. Bases were with respect to K113 for internal primers and from UCSC Genome Browser for flanking primers.

Table 2.2. Arrangement and sequences of primers for sequencing the 12q13.2 provirus.

Designation	Direction	Primer sequence (5' → 3')	Base position [a]
LTR6F	Forward	CTGAGTTGACACAGCACACG	173
LTR2F	Forward	CTGTGCTGAGGAGGATTAGT	433
LTR7F	Forward	TCCATATGCTGAACGCTGGT	830
LTR6R	Reverse	CGTGTGCTGTGTCAACTCAG	8677
LTR7R	Reverse	CCTTGGACAATACCTGGCTT	8839
LTR8R	Reverse	CCAGCCTCTGAGTTCCCTTA	9324
1F	Forward	TCTCTAGGGTGAAGGTACGC	993
2F	Forward	ATGTAGCAGAGCCGGTAATG	1511
3F	Forward	CCACAGTTGAGGCCAGATAC	2013
4F	Forward	AAGCCGGTAAGGTCATAGTG	2526
5F	Forward	CAGCCATTTGTTCCTCAGGG	2990
6F	Forward	GAAGGGTTGGTAGACACTGG	3511
7F	Forward	GTAAATCAGTGGCCGCTACC	3984
8F[b]	Forward	AGAGGTTGCCAATGCTGGAC	4532
9F	Forward	GGTCATTCCTTCCTCACAGT	4960
10F	Forward	ACAAGGGATGTTGAGACAGC	5451
12F	Forward	GCAACTTGCCAAACAGGAGA	5913
13F	Forward	TGTCCAAGTGCACAAGTGAG	7294
14F	Forward	TGGGAGGCCTCACCATCCGT	7777
15F[b]	Forward	AGGAGTTGCTGATGGCCTCG	8295
26F[b]	Forward	CCACTCCTCAGATGCAACTT	6202
27F	Forward	GAAGTACCTACTGTCAGTCC	6994
16R[b]	Reverse	TTGTTCTGGAAACCATGGGC	1266
17R	Reverse	ATACTGAAGTTCAGCCAGCG	3983
16R	Reverse	ATACTGAAGTTCAGCCAGCG	1266
18R	Reverse	TTAACAGGAGGATTGGCAGC	2276
19R	Reverse	TAAGTCAGGTGGCTCTCTAC	2853
20R	Reverse	GGCACTCCAAGGAATTGAAG	3329
21R[b]	Reverse	GAGCGGGCATGGTGATTTCC	3771
22R	Reverse	TTTCACAATCCTGCTCTGCC	4292
23R	Reverse	GTAGGAATGCCTAGAGTTGG	4734
24R[b]	Reverse	AAGACCAATCTGTTATGCAC	5149
25R	Reverse	GCCTGTTTCCATGTGACATC	5727
27R	Reverse	GGACTGACAGTAGGTACTTC	7012
28R	Reverse	GGTGTGAGGCCACAGTAAGC	7491
29R	Reverse	GTCTCCCATCCAAATGACAG	8063
30R	Reverse	TCTCTCTTGCTTTTCCCCAC	8524

[a] Relative to K113, with provirus start as position 1. Positions were mapped to the K113 nucleotide sequence using the NEBCutter v.2.0 web interface.

[b] Used in PCR amplification reaction paired with appropriate primers flanking the 12q13.2 locus: 12q13.2 F was paired with the above R; 12q13.2 R with above F.

Phylogenetic analyses

Alignments were generated from the full-length proviral LTRs using ClustalW with penalties set at 25 to for opening a gap, and 0 for gap extension. The resulting alignment was manually edited using BioEdit v.7.0.9.0, as for the full-length proviral alignment above. Elements no longer associated with either LTR were excluded from the analyses, however those with single LTRs were included here. The LTRs from all subgroups were included in the primary tree, and for subsequent trees the consensus LTRs of the remaining subgroups were included. Briefly, the 5' and 3' LTRs were aligned per subgroup and a consensus generated in BioEdit, with the most common base called for a given position. Individual LTRs from each subgroup were then aligned with the consensus of the other subgroups and trees generated with the same model and parameters as described here. Neighbor-joining (NJ) trees were generated in MEGA4 using the pairwise deletion option and a bootstrap value of 10,000 [176]. The Kimura 2-parameter (K2P) model was used for distance correction, with the α of the γ correction set to 1.5. Bootstrap values above 65 are provided for each tree. Maximum parsimony (MP) trees were generated with all sites included and tree options set at max-mini branch-and-bound to search all possible MP trees. Bootstrap was set at 10,000 replicates as above. For outgroup rooting to the initial tree, we attempted to, but were unable to align LTR sequences from other non-HML-2 HERV groups, and the tree was rooted at its midpoint due to this low similarity. For subgroup trees, the consensus of the most distant group consensus was used as the outgroup for the analyses; in support, branching patterns were not altered in MP trees rooted similarly.

For trees based on internal reading frames, only those elements with the region intact were included in the appropriate tree. For these purposes, the internal reading frames corresponding to *gag* (bases 1083-2923 with respect to K113), *pol* (bases 4986-5650) and *env* (bases 8047-9003) were extracted and aligned in ClustalW as described above. For analysis of the *pol* reading frame, we were able to align the corresponding regions from the canonical beta-like HERVs HML-1 through HML-10, in addition to the exogenous *Betaretroviruses*. Full length sequences were from HML-1 (AF015999), HML-3 (AF015998), HML-4 (U35160), HML-5 (AF015995), HML-6 (AF015997), HML-7 (AF016000), HML-8 (AF015996), HML-9 (AF016001), HML-10 (U07856), MMTV (NC001503), JSRV (M80216), and MPMV (NC001550). Maximum parsimony trees were generated as above. All tree topologies were confirmed using Bayesian inference (MrBayes v.3.1.2) [177, 178] with four independent chains run for 2,000,000 generations until sufficient trees were sampled to generate >99% credibility (Bayesian inference was with R.P. Subramanian).

Human DNA samples

Human genomic DNA samples were from one of two sources. The first was provided from the American Cancer Society (ACS) Cancer Prevention Study II Nutrition Cohort (CPSII) [179]. Genomic DNAs were isolated by the ACS from cheek swabs given by CPSII cohort participants. Original samples were 100 total: 50 samples were from participants with later reported breast cancer, and controls (n=25 per group); 50 samples were from participants with later reported prostate cancer, and controls (also 25 per group). Controls were from participants with no disease history who had not been

reported with the appropriate disease at the time of sample donation or release. Samples were blinded, and the keycode released following initial PCR analyses. Knowledge of the subjects was limited to case/control assignment (i.e., no information concerning disease status or type, or ethnicity of patients was disclosed). For statistical significance, secondary PCR screens were performed from a separate blinded sample set of 200 total genomic DNA samples (n=100 per breast cancer cases or controls), also provided from the ACS CPSII Nutrition Cohort. As above, all samples were blinded, and the keycode released following PCR analyses.

A second set of genomic DNA samples was from the Stanley Medical Research Institute (SMRI) from the Array Collection for research on schizophrenia and bipolar disorder [175]. Genomic DNAs were isolated from frozen sections of the occipital lobes from deceased participants. A total collection of 105 genomic DNAs were provided in 1μg quantities per sample. Samples were represented equally (n=35 per group) from deceased subject represented by clinically diagnosed schizophrenics, diagnosed cases of bipolar disorder, and undiagnosed controls matched for mean age, race, sex, in addition to other demographic features. The collection was predominantly from Caucasian subjects, and of 105 samples, just 2 were of other ethnicities (1 African-American, and 1 Native American within the bipolar group). As above, samples were provided as blinded and the keycode released following their initial analysis. In contrast to the samples provided from the ACS, subject demographics and histories were disclosed following unblinding of the sample set.

Whole genome amplification (WGA)

We included a whole genome amplification (WGA) (MIDI Repli-G, Qiagen) of all CPSII DNA samples prior to unblotting and PCR analyses to accommodate limiting amounts of DNA (~1µg per sample). WGA was carried out according the manufacturer's protocol with a starting volume of 5µL. Briefly, ~40ng genomic DNA per sample was denatured and neutralized using the supplied buffers in volumes of 5 and 10µL, respectively, each for 3 minutes at RT. A mixture containing buffered φ29 polymerase and random hexamers was added to each sample for a final volume of 50µL and the samples incubated 16 hr. at 30ºC for branching amplification of the DNA. Amplified DNA was extracted using heavy phase-lock gel (5PRIME) with 200µL volumes of phenol:chloroform:isoamyl alcohol (24:24:1) and chloroform:isoamyl alcohol (24:1). The aqueous phase was separated by centrifugation at 14,000 rpm for 5 min., and then transferred to 95% ethanol + 100mM sodium acetate, pH 5.2 to a final volume of 1mL and precipitated overnight at -20ºC. The WGA DNA was pelleted at 14,000 rpm for 30 min. at 4ºC, washed in 1mL 70% ethanol, and the centrifugation repeated. The ethanol was aspirated from each pellet, which was dried 30 min. at 37ºC and resuspended in 100µL sterile water.

PCR for screening of polymorphic HML-2 proviruses

For a total of 11 loci with evidence of multiple alleles that included the provirus form, primers were designed to amplify the 5' LTR of the provirus at each site using the published human sequence (GRCh37/hg19). For each locus, a primer was designed within ~1kb from the start of the provirus in the flanking DNA of the host, and

the second primer internal to, but near, the 5' LTR. A third primer was designed in the host regions opposite the integration site to amplify the remaining alleles. Primers were designed using Primer3 v.0.4.0 and from IDT, unless otherwise noted in the table legend. *In silico* PCR (UCSC Genome Browser) was used as an initial test of target amplification and product size, as provided in Table 2.3. All PCRs were carried out using ~200ng WGA DNA (also refer above for WGA) as template in a *Taq*-based amplification with 1.5-3.5 μM Mg^{++}, 200μM dNTPs, 0.2 μM each primer, and 2.5 U Platinum *Taq* (Invitrogen). Reactions were run under cycling conditions using a BioRad C1000 thermocycler as follows: Initial denaturation and *Taq* activation was at 95°C for 3 min, followed by 35 cycles of 95°C for 30 sec, primer annealing at a T_m of 56°C-58°C for 1 min, and an extension at 72°C for 1-2.5 min, based on the size of the expected amplification product with ~1 min per kb. Final extension was at the same temperature for 10 min. 10uL of each PCR reaction was analyzed by electrophoresis through 1% agarose in 1 x TBE. Products from 2 separate positive PCRs per primer set were sequenced from both directions to confirm the desired product and rule out nonspecific or background amplification.

Table 2.3.Primers and PCR conditions for detection of polymorphic HML-2 proviruses.

Locus [a]	Forward (5'→3')	Reverse (5'→3')	$MgCl_2$ [mM]	size (bp) [b] solo/pre
1p31.1 I	AACTACGTGAAGAATGAAGA	AATAAAGCTGAGATAAGAGG	3.5	1239
3q13.2 I	GCTCGGATTTCAACATCCAT	TCGTCCGACTTGTCCTCAATG	2.5	1821
3q13.2 II	GCTCGGATTTCAACATCCAT	TATTGGTGACAGAGAGATGCAG	2.5	1847/879
6q14.1 I	TCGTCGACTTGTCCTCAATG	CTGCCAGTCTCAGGTGTTTG	1.5	1075
6q14.1 II	CCCCTGCTTATTGATGCTCTACG	TGAGGCTGAATGTGTGGAGTCC	1.5	1526/556
7p22.1a I	TACTGAACGATGCTGACGTTTGG	TTTGAACCATTATCACCCTA	2.5	1407
7p22.1b T	GTCTGCAGGTGTACCCAACAG	TTTGCCCCATTATCACCCTA	2.5	1216
7p22.1 II	CCTCCTGGTTCAAGGGATTCTC	GCTTTCGGGACTTCAACATTGG	1.5	1387/419
8p23.1a I	CTTGTGTTTTCATTACAATCTATT	TTCAGTCATTCTATCATTAAGATTC	1.5	1667
8p23.1a II	CAGTCTATAGATGTGGATGCCT	AGCACTGAATCCAAACTCATAT	1.5	1320/352
10p12.1 I	CCACCATCTGAGAAGTGTGATG	AATGGAGTCTCCYATGTCTACT	2.5	1342
10p12.1 II	CCACCATCTGAGAAGTGTGATG	GGCAACAAAGGGTTCATATGAGAA	2.5	1508/540
11q22.1 I	CCATGCTCAGAAAGGAAACA	TAGCTTCTTCCGAGCACACA	2.5	1168
11q22.1 II	CCATGCTCAGAAAGGAAACA	ACCATCTGTCCTTCCACCAG	2.5	1661/693
12q13.2 I	CGGAGAATTCCACCTTCAAA	CTCGAGCGTACCTTCACCCTAG	2.5	1377
12q13.2 II	CGGAGAATTCCACCTTCAAA	TGCATTGTGGTCATCCATTT	2.5	1488/520
12q14.1 I	GGAAACCCTTCCAACATTCCA	CCCCATTATCACCCTAGCTTC	2.5	1299
12q14.1 II	GGAAACCCTTCCAACATTCCA	TGAGGCTGAATGTGTGGAGTCC	2.5	1101/133
19p12b I	TGCATGGGGAGATTCAGAACC	TCGGGATCTCTCGTCGACTTGTCC	2.5	1210
19p12b II	TGCATGGGGAGATTCAGAACC	CGTGTTAGCCAGGATGGTCT	2.5	310/1278

[a] For each tested site: set I primers were for the detection of the 5'LTR; set II primers were for the detection of either the solo LTR or empty site.
[b] All product sizes were estimated based on *in silico* PCR of primer sets I and II. Product sizes for the amplification of each allele for the 10p12.1, 12q13.2, and 19p12b proviruses were estimated manually, with reference to the primer site in the host genome regions flanking the individual integration sites.

Statistical analyses

All case-control comparisons were analyzed by χ^2 with one degree of freedom. For these analyses, comparisons were only between cases and corresponding controls per sample group, with 50 total samples for the ACS sample set (25 breast cancer cases and controls), and 70 total samples for the SMRI sample set (35 each schizophrenic cases and matched controls). All other samples were excluded from statistical analysis but were included in preliminary PCR screening. A *p* value of less than 0.05 was taken to be significant. Secondary PCR screening was performed from a separate blinded sample set from the ACS CPSII Nutrition Cohort. Total numbers of samples for screening were determined based on the individual provirus frequencies per group from primary PCR screening. For example, in the case of K115, for a 20% difference with an α=0.5, for a statistical level of 80% power, total sample size was calculated at n=94 (47 per cases and controls), and for 90% power a total sample size of n=124 (62 each cases and control). For these purposes, an additional 200 genomic DNAs were screened by PCR and the results analyzed in the same manner. All statistical analyses were outsourced to the Data Design and Resource Center at Tufts University.

***In silico* restriction analyses**

We used an *in silico* approach to identify useful restriction enzymes for subsequent DNA blots to visualize HML-2 proviruses, and to generate a comparison from sequenced genome data for reference in unblotting (described below). In a similar strategy to that used in exhaustive searches for HML-2 as above, we mined the most recent genome build (GRCh37/hg19, February 2009) by BLAT search for sequences with

high percent identity to K113. The cellular sequence flanking each provirus returned from the search was examined to confirm that the full-length of the element was identified, and to verify any near full-length or truncated elements. Complete length nucleotide sequences were generated per site by extracting and concatenating the internal and LTR segments, for 62 elements. As described above, full-length sequences for four additional HML-2 were from NCBI, at 10p12.1 (AF164611), 19p12b (AY037928), K105 (AC195095.), and the provirus at 12q13.2 that was sequenced presently (R.P.S., J.H.W., *et al*., in preparation). Secondary searches were not performed with the internal reading frames from K113, and those elements were not included in the analysis. Close examination of the alignment revealed an HML-2-specific region spanning bases 1017 to 1049 (5' CGTCGACTTGTCCTCAATGACCACGCTCGAGC 3'). We verified the specificity of the 32 bp sequence by BLAT-searching the Hg19 genome build. A total of 34 proviruses corresponding to the HML-2 with highest sequence similarity to K113 were initially hit, from which 25 hits were identified that had two or fewer mismatches and 17 were 100% identical. This sequence, which we refer to as 'K-seq', was utilized as an indicator of HML-2 elements in further *in silico* analysis and for subsequent unblotting.

In silico restriction analysis was performed as follows. The HERV-K113 (AY037928) sequence was analyzed in NEBCutter2.0 for restriction enzymes predicted to cut at least once within the provirus but not within the 5'LTR. For each of the 19 proviruses with the HML-2-specific K-seq site intact, we performed an *in silico* restriction analysis as follows. About 5kb of sequence was extracted in either direction from the start of the 5'LTR within the UCSC Genome Browser. Each sequence was

analyzed in NEBCutter v.2.0 in an '*in silico* digest' using each of the commercially available restriction enzymes. The conserved K-seq site was utilized as a reference point for the closest identified restriction sites. In total, we analyzed the 5' LTR and flanking sequence of each provirus with all available restriction enzymes, of which 36 cut at least once in the provirus and flanking DNA, but not within or upstream of the K-seq site or in the 5' LTR. These 36 restriction enzymes were also selected for *in silico* analysis based on the criterion that we excluded enzymes that recognized site either >6 bp or <4 bp, that cut at sites with multiply ambiguous bases, or that cut distal to the site.

In our hands, the use of WGA-DNA as the template for DNA hybridization in unblotting carried an inherent size limitation, such that fragments larger than ~5-6kb in length were not readily detected. For each potential enzyme, we performed an *in silico* analysis to generate the predicted product size ranges and fragment distributions. Per enzyme, the size estimates for predicted HERV-K-containing junctions fragments were plotted on a log scale for comparison of size range and product distribution. Six enzymes were selected for analysis of WGA-DNA by unblot (described below); two enzymes, *Bsr*I and *Dra*I, were found to each generate a well-distributed fragment and reproducible pattern ranging in size from ~1-6kb. Both were selected for preliminary unblot analysis and coordinate *in silico* comparison; subsequent unblot analyses were with *Bsr*I.

Unblotting

Unblotting, or hybridization in semi-dried agarose [180], was carried out to visualize polymorphic HERV-K proviruses within CPSII DNA samples. For each sample, 15µg of WGA DNA was digested with *Bsr*I (New England Biolabs) in a 100µL volume and the digested products extracted and precipitated as described above. Products were resuspended in 20uL 0.25 x TBE + 30% ficol and electrophoresed through a 0.8% agarose gel in 0.25 x TBE at 70V for 29 hr. at 4ºC. The gel was dehydrated in a gel dryer (BioRad) layered on filter papers and Saran wrap for 60 minutes at RT and 60 min. at 62ºC. The dried gel was stained with ethidium bromide and excess agarose removed with a clean scalpel. The gel was then incubated in denaturing buffer (0.5M NaOH + 1.5M NaCl), and neutralizing buffer (1.0M Tris-HCl + 1.5M NaCl, pH 8.0) 30 min. each at room temperature, and then hybridized 7.5 x 10^6 cpm of the ^{32}P-labeled conserved HML-2 oligonucleotide at positions 1017-1049 of K113 (5' CGTCGACTTCTTGTCCTCAATG ACCACGC). The oligonucleotide probe was labeled in a T4 polynucleotide kinase reaction (New England Biolabs) according to the manufacturer's suggested protocol with 5µM ATP[γ^{32}P] (Perkin Elmer). Hybridization was in 5x SSPE + 0.1% SDS, pH 7.4 at 53ºC for 16 hr with shaking at 50 rpm. Following hybridization, the gel was washed (2x SSC + 0.1% SDS) 4x for 15 min. each at room temperature, and 2x for 30 min. each at 53ºC with shaking at 70 rpm. The gel was then exposed to BioMax MS film (Kodak) under an intensifying screen for 4-5 days at -70ºC before developing.

Chapter 3

Results and Discussion Part 1:

Identification, Characterization, and Comparative Analysis of the HML-2 Group of Human Endogenous Retroviruses

Significance

Since its publication in 2001 [2], the available draft of the human genome sequence has evolved through several builds, through which defined genome coordinates have been altered. These changes, although necessary, present a problem in confirming the chromosomal locations and exact copies of referenced sequences such as various REs, and have complicated the use of existing literature for the verification of individual HML-2 loci. In part, the fact that members of HML-2 are polymorphic has led to an incomplete representation in the existing genome sequence. For example, the K103 provirus (located at chromosomal position 10p12.1) is represented as a solo LTR in all genome builds, however the provirus has been sequenced and is publicly available in the NCBI nucleotide database [134, 181]. Also missing from the published sequence is HERV-K113, located at 19p12 and arguably the most studied and well-characterized HML-2 provirus. It is likely new polymorphic integrations of HML-2 HERVs will be identified in the future, would be of interest to more than a few areas of research. The absence in current genome builds of some HML-2 gives an already incomplete representation of the group, and it is a concern that newly identified elements may be excluded as well. Knowledge of the existing elements –or at least their indication and eased reference within such databases– would decrease complications in future studies, and during the prospective discoveries of new sites.

Here, we report a comprehensive analysis of HML-2 elements present within human DNAs. Through iterative sequence mining of the most recent human genome assembly (Feb.2009 GRCh37/hg19), we have identified and characterized 89 full-length, near full-length, and partially degenerated proviruses belonging to the HML-2 group of

HERVs. We have accounted for all known polymorphic HML-2 proviruses, including those within chromosomal positions 10p12.1 (K103), 19p12b (K113), and the K105 provirus that is located within an unaligned contig. We have also sequenced and included the previously uncharacterized HML-2 provirus at the 12q13.2 locus, also represented in the published genome as a solo LTR. Using this current and most complete 'catalog' of HML-2, we have analyzed this group of HERVs to determine their intragroup relationships and evolutionary patterns. We have verified and further characterized specific subgroups within the HML-2 group, and report a high frequency of the type 1 elements within the most recently active and successful HML-2. Finally, we present the characterization of three identified sub-clades within HML-2 that have been amplified as the result of segmental duplications, with the conspicuous association of additional HML-2 provirus to a particular segment type. We discuss these observations and speculate their implication in future genome evolution.

Generation of a comprehensive dataset of the HML-2 group

We mined the most recent human genome assembly (GRCh37/hg19) for sequences with strong similarity to HERV-K113 using the BLAST alignment-like tool (BLAT) within the UCSC Genome Browser website [169]. The K113 provirus (AY037928), located at chromosomal band 19p12, is reported to be the most recent retroviral integration described in the *Homo* germline and has retained complete ORFs. The K113 provirus has a detectable allele frequency of ~16%, and has been estimated to have formed ~100,000-200,000 years ago, based on the nucleotide differences between its 5' and 3' LTR, and normalized to a neutral substitution rate of 0.2% per million years [95, 182]. Given its relatively recent activity, insertional polymorphism, and high level of conservation, K113 has been one of the most studied proviruses in terms of infectious capacity, function, and possible correlation to disease (reviewed in [76, 183]). By searching for proviruses with the highest percent identity to K113, we were able to identify 62 full-length or near full-length proviruses with >87% nucleotide identity to the full-length K113 genome (Table 3.1). Additional searches using nucleotide sequences for the consensus HML-2, HERV-K_{CON} and K-PHOENIX, both >99% identity to K113, resulted in the same set of identified sites, supporting our search strategy.

We added an additional four proviruses that were reported from the literature: K113 itself, K103 (located within the 10p12.1 band), K105 (located within an unassembled centromeric region Un_g1000219), and an 'unnamed' provirus located at 12q13.2. The insertion site for K113 is represented in all published genome builds as an empty, or pre-integration site, whereas the latter three sites have been published as solo LTRs. The K113, K103, and K105 proviruses were previously identified in HML-2-

specific DNA hybridizations of human-derived genomic BAC libraries [95, 173] and for these sites the full-length sequences available from GenBank were added to our initial dataset (K113, AY037928; K103, AF164611; and K105, GU476554) (Table 3.1).

As mentioned, the insertion at 12q13.2 was previously found to be polymorphic by Belshaw *et al*. in a large-scale PCR-based screen designed to detect polymorphic alleles of human-specific solo LTRs [131]. Previous analysis of the 12q13.2 integration site confirmed the presence of all three alleles: provirus, solo LTR, and pre-integration [131]. Using a set of conserved primers, we amplified and sequenced the full-length 12q13.2 provirus (R.P.S., J.H.W., *et al*., in preparation; also described further below). The 12q13.2 site represents one of the few 'new' polymorphic HML-2 to be made available since the human genome publication, and its addition here presents 66 HML-2 identified within our dataset with high similarity to the full-length K113 (Table 3.1).

We expanded our original search to detect HML-2 proviruses with lower percent identity to individual HML-2 reading frames but which still belonged to the HML-2 group. For these added searches, we used the criteria that a provirus should have at least ~50% of the internal portion being present and/or the association of an LTR in the same orientation and within 2kb. In subsequent individual searches, we BLATed the *gag*, *pro*, *pol*, and *env* reading frames with reference to K113, to the 2009 human genome build. Each search confirmed our initial BLAT hits from the full-length K113 sequence, but also led to the identification of 17 additional elements (asterisked in Table 3.1). As above, repeated searches with the consensus ORFs belonging to the engineered HML-2 confirmed and supported these results (data not shown).

A final search was performed with reference to the UCSC RepeatMasker. The UCSC Genome Browser utilizes an automated classification system for stretches of repetitive elements (REs), referred to as RepeatMasker, which assigns each element according to class and includes those derived from the DNA (i.e., Tigger), LINE (i.e., L1), SINE (i.e., *Alu*), LTR (i.e., HERV-K), or 'other' (i.e., microsatellite) classes. Working from the 2009 genome build, we filtered all repetitive DNA using RepeatMasker for those derived from HML-2, which are classified as either 'HERV-K-int' (internal proviral sequence) or belonging to the LTR subgroups represented within HML-2 (LTR5-Hs, LTR5A, or LTR5B [94, 115]). This led to the tentative identification of an additional 8 elements that fit the pre-established criteria for a provirus as described above. However, in subsequent analyses (below), only 6 of the 8 elements were observed to group most closely with HML-2, and so the remaining two elements were excluded from the final HML-2 dataset. With the addition of the 6 elements (asterisked in Table 3.1), a total of 89 HML-2 insertions was confirmed to exist within humans.

Overall, we have generated an exhaustive dataset of the full length and near full-length HML-2 proviruses with evidence for the provirus in humans. To the best of our knowledge, this represents the most detailed catalogue of HML-2 elements to date, adding >30 proviruses to the most recent report [120], and including ~27 proviruses that were previously undescribed (also indicated in Tables 1 and 2) [64, 94, 120, 173, 184-188]. Using this dataset, we characterized the HML-2 group of HERV-K with reference to LTR5 subgroup (Hs, A, or B), subtype (type 1 or 2), and distribution throughout the genome (R.P.S., J.H.W., *et al*., in preparation). The results are detailed in the following sections.

Table 3.1. Collection of the HERV-K (HML-2) group used in this study.

locus	alias	start (bp)	age[d]	# dif. LTRs[d]	LTR	reference
[a] 1p31.1	K4	75842771	<2	0	Hs	[89]
*1p34.3		36954585			Hs	this report
*1p36.21a		12840258			B	this report
1p36.21b	K(OLD-AL023753), K6, K76	13458305	18.80	97	B	[90]
*1p36.21c	K6, K76	13679141	18.99	98	B	[120]
1q21.3		150605284			Hs	this report
1q22	K102, K(C1b), K50a	155596457	<2	2	Hs	[173]
1q23.3	K110, K18, K(C1a)	160660575	6.59	34	Hs	[173]
1q24.1	K12	166574603	13.76	71	B	[120]
1q32.2		207808457			Hs	this report
*1q43		238925595			B	this report
2q21.1		130719538			Hs	this report
3p12.3		75600465			A	this report
3p25.3	K11	9889346	12.79	66	Hs	[64]
3q12.3	K(II)	101410737	5.23	27	Hs	[188]
[a] 3q13.2	K106, K(C3), K68	112743479	<2	0	Hs	[173]
3q21.2	K(I)	125609302	3.48	18	Hs	[188]
3q24		148281477				JFH, JMC, unpub.
3q27.2	K50b	185280336	<2	3	Hs	[89]
*4p16.1a	K17b	9123515	13.57	70	A	[120]
*4p16.1b	K50c	9659580	14.53	75	A	[120]
4p16.3a		234989			B	this report
*4p16.3b	K77	3979051	10.08	52	A	[120]
4q13.2		69463709	13.95	72	A	this report
4q32.1		161579938			Hs	[89]
4q32.3	K5	165916840	7.75	40	Hs	[89]
4q35.2		191027414	10.46	54	A	this report
5p12		46000159	10.66		Hs	[120]
5p13.3	K104, K50d	30487114	5.23	27	Hs	[173]
5q33.3	K107, K10, K(C5)	156084717	<2	2	Hs	[118]
*5q33.2	K18b	154015513	11.43	59	A	[120]
*6p11.2	K23	57622734	8.33	43	A	[120]

locus	alias	start (bp)	age[d]	# dif. LTRs[d]	LTR	reference
6p21.1	K(OLD-AL035587),	42861409	8.52	44	B	[90]
6p22.1	K(OLD-AL121932), K69, K20	28650367	7.56	39	Hs	[90]
[a] 6q14.1	K109, K(C6)	78427019	<2	3	Hs	[173]
6q25.1		151180749			B	this report
[a] 7p22.1a	K108L, HML2-HOM, K(C7)	4622057	<2	2	Hs	[90]
[a] 7p22.1b	K108R	4630561			Hs	[90]
*7q11.21		65469671			B	this report
7q22.2		104388369			Hs	this report
7q34	K(OLD-AC004979)	141450926			Hs	[90]
[a] 8p23.1a	K115	7355397	4.07	21	Hs	[95]
8p23.1b	K7	8054700	13.97	72	A	[89]
8p23.1c		12073970	13.57	69	A	[89]
8p23.1d	K(OLD130352)	12316492	13.95	70	A	[89]
8q11.1	K70, K43	47175650	19.77	102	A	[120]
8q24.3a		140472149		52	Hs	this report
*8q24.3b	K29	146246648	10.08		A	[64]
9q34.11	K31	131612515	10.85	56	B	[64]
9q34.3	K30	139674766	13.18	68	B	[64]
[a,b] 10p12.1	K103, K(C10)	27182399	<2	7	Hs	[173]
10p14	K(C11a), K33	6867109	6.20	32	Hs	[89]
10q24.2		101580569			Hs	[94]
11p15.4	K7	3468656	13.18	68	A	[120]
11q12.1		58767448				this report
11q12.3	K(OLD-AC004127)	62135963	14.73	76	Hs	[90]
[a] 11q22.1	K(C11c), K36	101565794	<2	4	Hs	[89]
11q23.3	K(C11b), K37	118591724	15.9	55	Hs	[186]
12p11.1	K50c	34772555	31.59	163	A	[120]
[a,c] 12q13.2		55727215	<2	1	Hs	[131]
[a] 12q14.1	K(C12), K41	58721242	<2	4	Hs	[186]
12q24.11		111007843			Hs	[67]
*12q24.33	K42	133667120	6.20	32	A	[120]
*14q11.2	K(OLD-AL136419), K71	24480625				[90]
*14q32.33		106139650			A	[120]

locus	alias	start (bp)	age[d]	# dif. LTRs[d]	LTR	reference
*15q25.2		84829020				this report
16p11.2		34231474			Hs	this report
*16p13.1	K(OLD-AC004034)	2976160				this report
19p12a	K52	20387400	22.68	117	B	[64]
[a,b] 19p12b	K113	21841536	<2	3	Hs	[95]
19p12c	K51	22757824	10.46	54	Hs	[120]
19p13.3		385095			Hs	this report
19q11	K(C19)	28128498			Hs	[187]
19q13.12a		36063207			Hs	this report
19q13.12b	K(OLD-AC012309)	37597549	17.83	92	Hs	[90]
*19q13.41		53243693			B	[64]
19q13.42	LTR13	53862348			B	[64]
20q11.22	K(OLD-AL136419), K59	32714750	11.04	57	B	[64]
21q21.1	K60	19933916	<2	5	Hs	[84]
22q11.21	K101, K(C22)	18926187	<2	8	Hs	[173]
*22q11.23	K(OLD-AP000345),	23879927	6.39	33	B	[64]
[b,c] U219	K105	175210	9.88	38	Hs	[173]
*Xq11.1		61959549				this report
*Xq12		65684132				this report
Yq11.23a		26397837				this report
Yq11.23b		27561402				this report
*Xq28a	K63	153816922			B	this report
*Xq28b	K63	153836676	12.60	65	B	[94]
*Yp11.2		6826441			A	this report

* Detected in BLAT-searches of the Hg19 human genome build with individual reading frames from the K113 provirus.

[a] Described as polymorphic in humans.

[b] The 19p12b provirus, K113, is: not represented in the present genome build Hg19. The elements corresponding to 10p12.1, 12q13.2, and U219 are published as solo LTRs in the Hg19 build.

[c] 12q13.2, U219: this is the first report of the full-length sequence.

[d] Calculations were performed with the help of R.P. Subramanian (R.P.S., J.H.W., *et al.* in preparation). Times of formation estimated at less than 2 years are indicated as such; also refer to the test for details.

Sequence and analysis of a polymorphic HML-2 provirus located at 12q13.2

The insertion at 12q13.2 was previously found to be polymorphic by Belshaw *et al*. in a large-scale PCR-based screen designed to selectively amplify polymorphic alleles with reference to human-specific solo LTRs [131]. Analysis of the site at 12q13.2 confirmed not only the presence of the solo LTR, but of the provirus and the pre-integration site from a panel of ethnically diverse human DNA samples [131].

We verified the 12q13.2 insertion site, and sequenced the full-length provirus for inclusion in the present analyses (Figures 3.1 and 3.2). Briefly, we searched the Hg19 genome build with the published DNA sequence flanking the 12q13.2 solo LTR [131]. With reference to this site, locus-specific primers were designed to amplify either the solo LTR or pre-integration alleles (Figure 3.1 A). Pairing the appropriate flanking primer with an internal HML-2-specific primer allowed for amplification of the 5'LTR to indicate the presence of the proviral allele. Using the primers, we screened a set of ~35 genomic DNA samples in two PCR reactions to determine the allelic distribution for the 12q13.2 site.

The most common allele observed within the samples screened was the provirus, with an allele frequency of 0.527, followed by the solo LTR and empty sites, with detected frequencies at 0.306 and 0.167, respectively. Thus, over half of the occupied sites at the 12q13.2 locus were from proviral forms, as observed in 14 of 18 individuals with at least one copy (Figure 3.1 B). Overall, the alleles were detected at similar frequencies as originally reported, with the exception that the solo LTR was the least common allele, with more individuals homozygous for either the provirus or empty

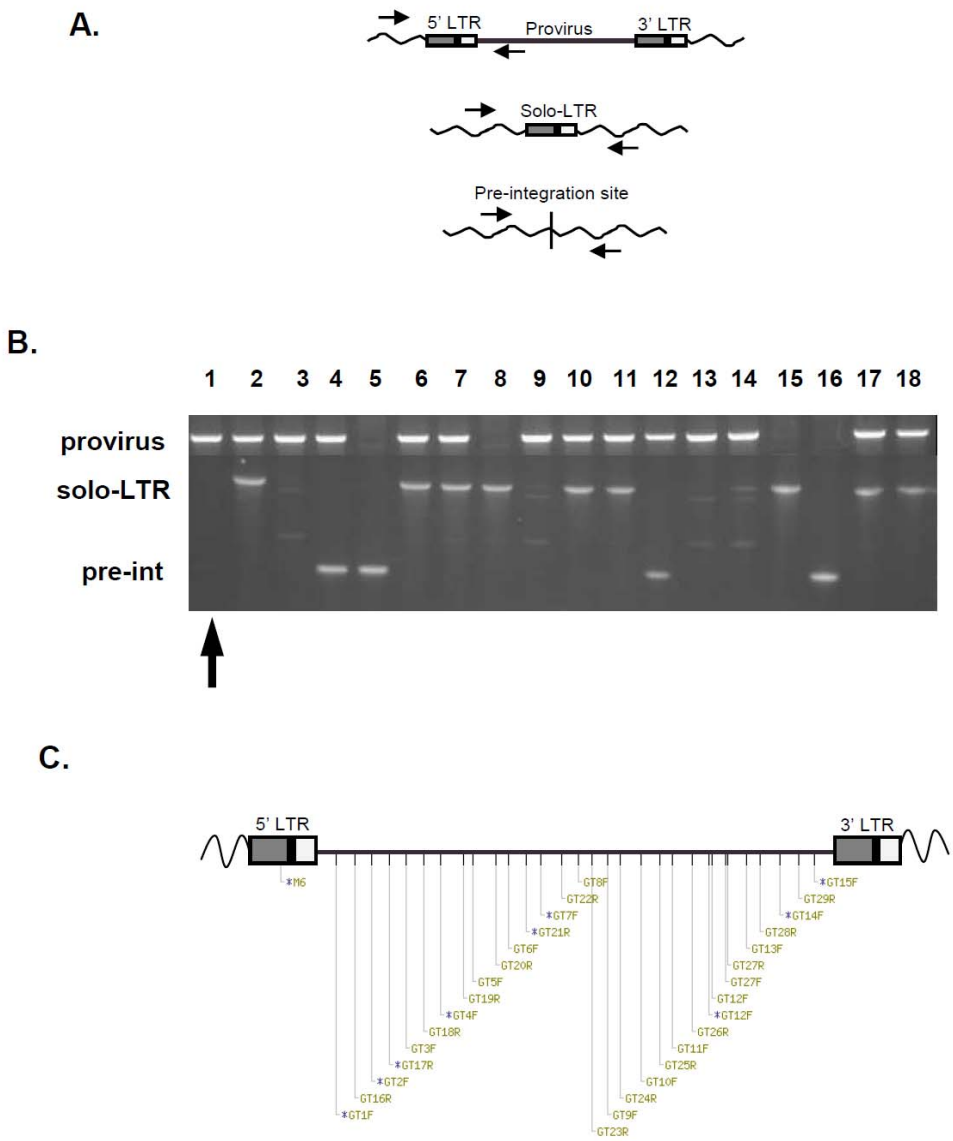


Figure 3.1. Strategy for the detection, amplification, and sequencing of the 12q13.2 HML-2 provirus.

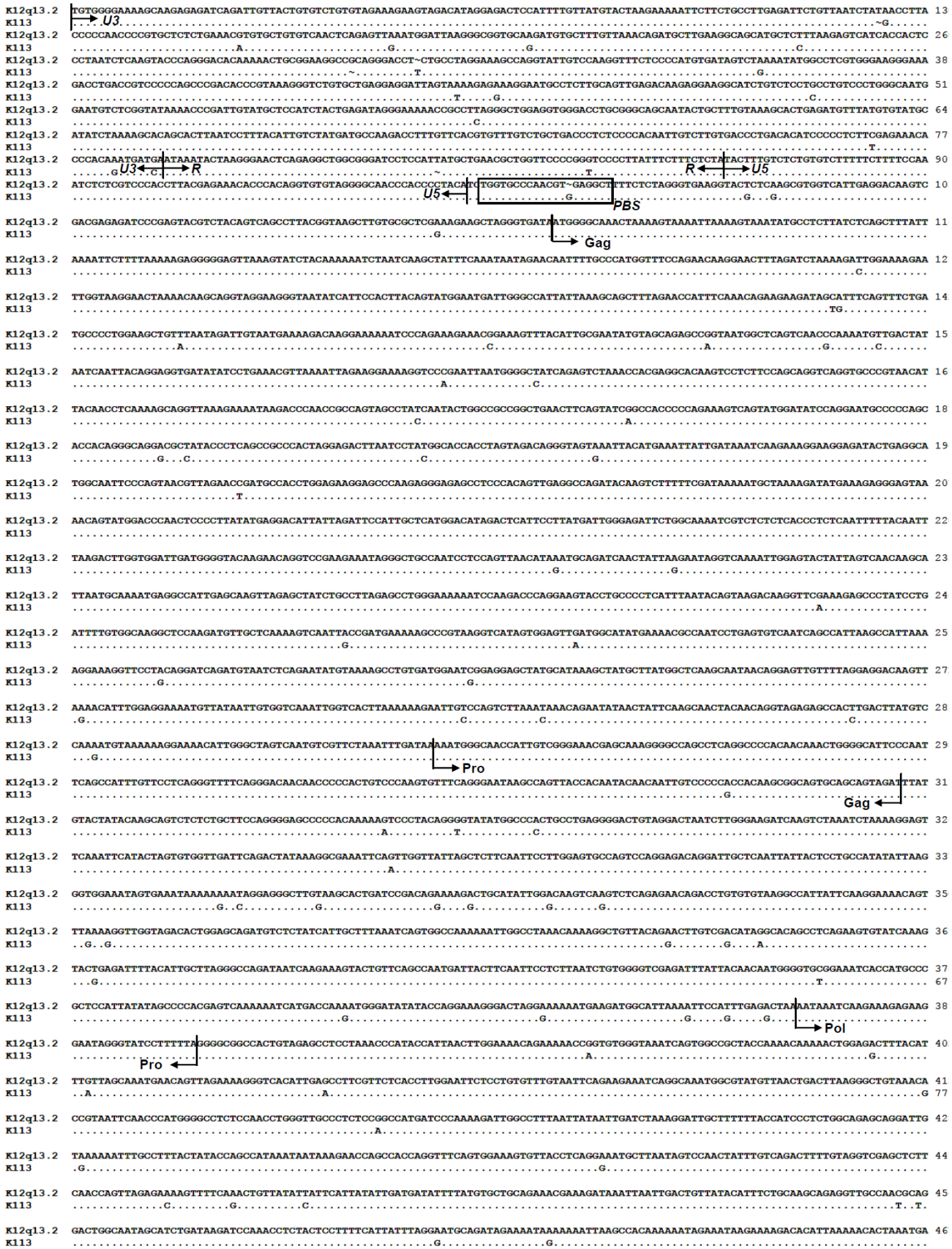
A. PCR strategy to detect the 5' LTR, solo LTR, or unoccupied pre-integration alleles for the 12q13.2 provirus. The 5' LTR is amplified with a primer flanking the insertion site, and to HML-2-specific sites within provirus. A second PCR detects the solo LTR or empty site; either allele is discriminated by a difference in size of the LTR (986 bp).

B. PCR results to detect the allelic distribution of the 12q13.2 site in human genomic DNAs. Arrow indicates an example of a DNA sample with two copies of the full-length provirus. This sample was used for amplification and sequencing of the full-length allele.

C. Sequencing strategy for the provirus using a set of primers conserved to the HML-2 group [173]. The indicated sample was amplified in two overlapping fragments and PCR products sequenced to ~6x coverage with the primers as indicated.

site (frequencies were 0.49 for provirus, 0.10 solo LTR, and 0.41 empty site) [131]. This discrepancy was likely due to the samples used for screening, for which we have little information for ethnicity, and lends a plausible explanation for the differences in our observed allele frequencies for 12q13.2 from the original report.

From a selected sample with two proviral copies, we amplified and sequenced the provirus to ~6x coverage using a set of conserved internal primers from an individual with two copies of the provirus present (no detectable solo LTR or empty site) (Figure 3.1 C). Following sequencing, a full-length consensus was derived and presented here (Figure 3.2). The 12q13.2 full-length nucleotide sequence was found to be fairly conserved, with 94.7% identity to K113 and 95.2% identity to the HML-2 consensus HERV-K_{CON}. The nucleotide sequence contained two inserted stop codons, one within the *pol* ORF at 12q13.2 position 5722, and the other within the start of *env* at position 6495 (labeled in Figure 3.2). Intact reading frames were present across the *gag* and *pro* ORFs. The provirus was found to be a type 1 HML-2, containing the specific deletion of 292bp at the *pol*-*env* division. The addition and primary characterization of the 12q13.2 provirus to this dataset represents one of several 'new' polymorphic HML-2 to be made available since the human genome publication (Table 3.1).


U3
U3
R
R
U5
U5
PBS
Gag
Pro
Pro
Gag
Pol

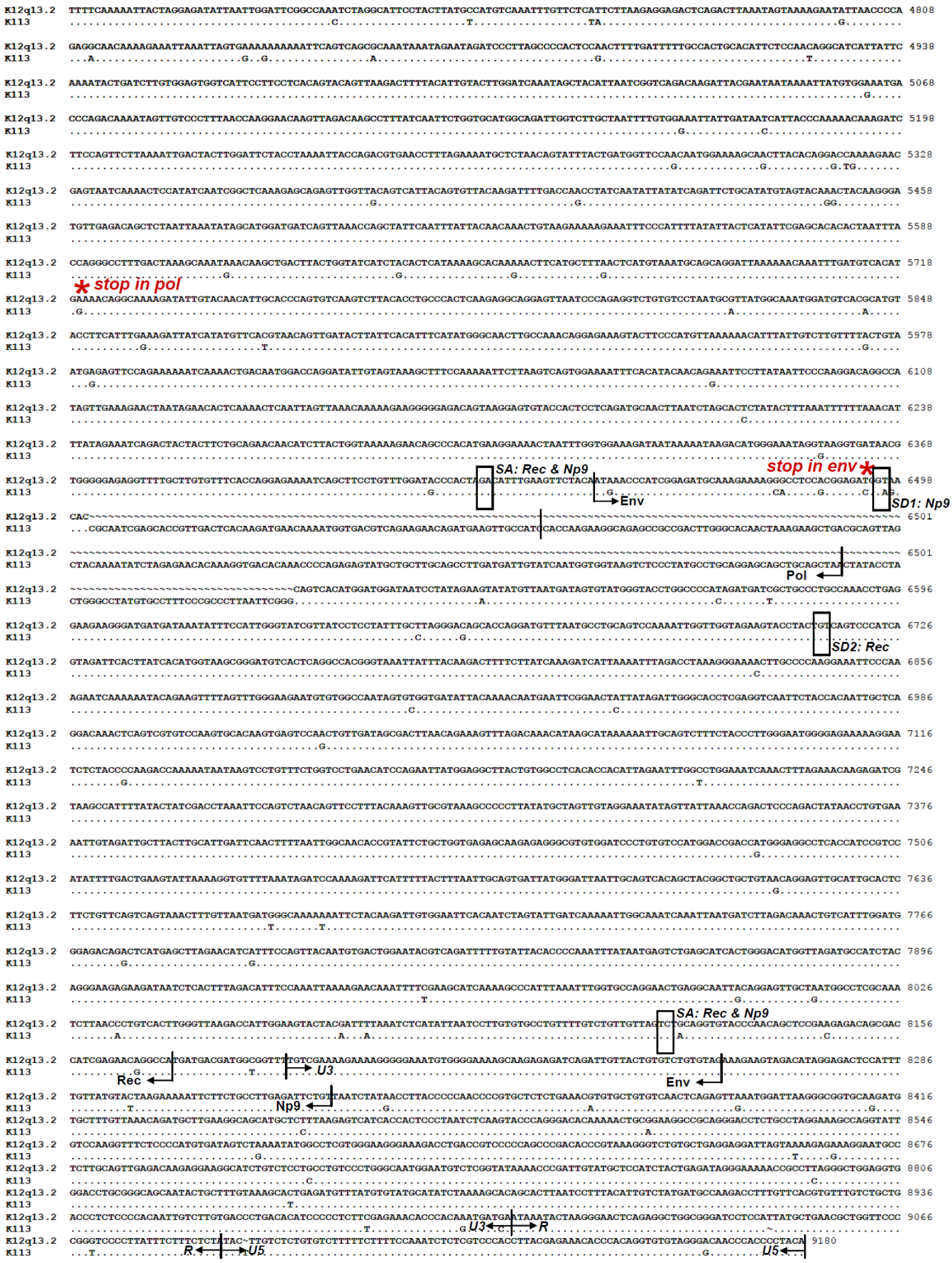

stop in pol
SA: Rec & Np9
Env
stop in env
SD1: Np9
Pol
SD2: Rec
SA: Rec & Np9
Rec
U3
Env
Np9
U3
R
R
U5
U5

Figure 3.2. Nucleotide sequence and annotation of the 12q13.2 HML-2 provirus.

The consensus sequence of the 12q13.2 type 1provirus aligned to the full-length K113. Features specific to HML-2 are indicated. Reading frames are each delineated and arrowed. Premature stop codons in *pol* and *env* are each marked with an asterisk. ORFs for the type 2 Rec and type 1 Np9 are marked in the same manner, and their respective splice donor (SD) and acceptor (SA) sites labeled and boxed, and were from Armbruester *et al*., in 2002 [122]. Boundaries for the U3, R, and U5 LTR regions are delineated for the inferred transcriptional start and stop signals, respectively at the U3-R, and R-U5 boundaries; these marked sites were recently reported as specific for the HML-2, and were from [130]. The primer binding site (PBS) is boxed downstream of the 5' LTR edge.

Analysis of the HML-2 group of proviruses in humans

As an initial analysis of the HML-2 proviruses, we characterized each element in terms of overall nucleotide sequence and structure with reference to K113. A nucleotide alignment was generated in ClustalW of all 89 sequences and manually edited using the BioEdit program. Using the aligned sequence information as a guide, the structures of individual were characterized to include the length, insertions and deletions, ORFs, and introduced stop codons. Based on these features, the inferred structure was mapped for each provirus, and is shown schematically in Figure 3.3. Each provirus has also been diagrammed and arranged based on specific subgroupings within the HML-2 group, as determined in this study (described in the following sections).

A third (~30 of 89) of the identified HML-2 proviruses were found in the full-length forms, including ~7 insertions missing some or all of a single LTR. As could be expected, many of the intact proviruses were human-specific elements, as previously determined by comparative PCR with humans and other primates (~20 were human-specific; single crossed bars in Figure 3.3), including those known to exhibit polymorphic alleles within humans (~10 of 11 full-length elements were polymorphic; double crossed bars in Figure 3.3). The majority of remaining insertions were degenerate and heavily mutated, with multiple stop codons and large deletions or insertions, an observation that was particularly true of those elements identified in the secondary searches using individual K113 reading frames, or extracted with reference to RepeatMasker classification.

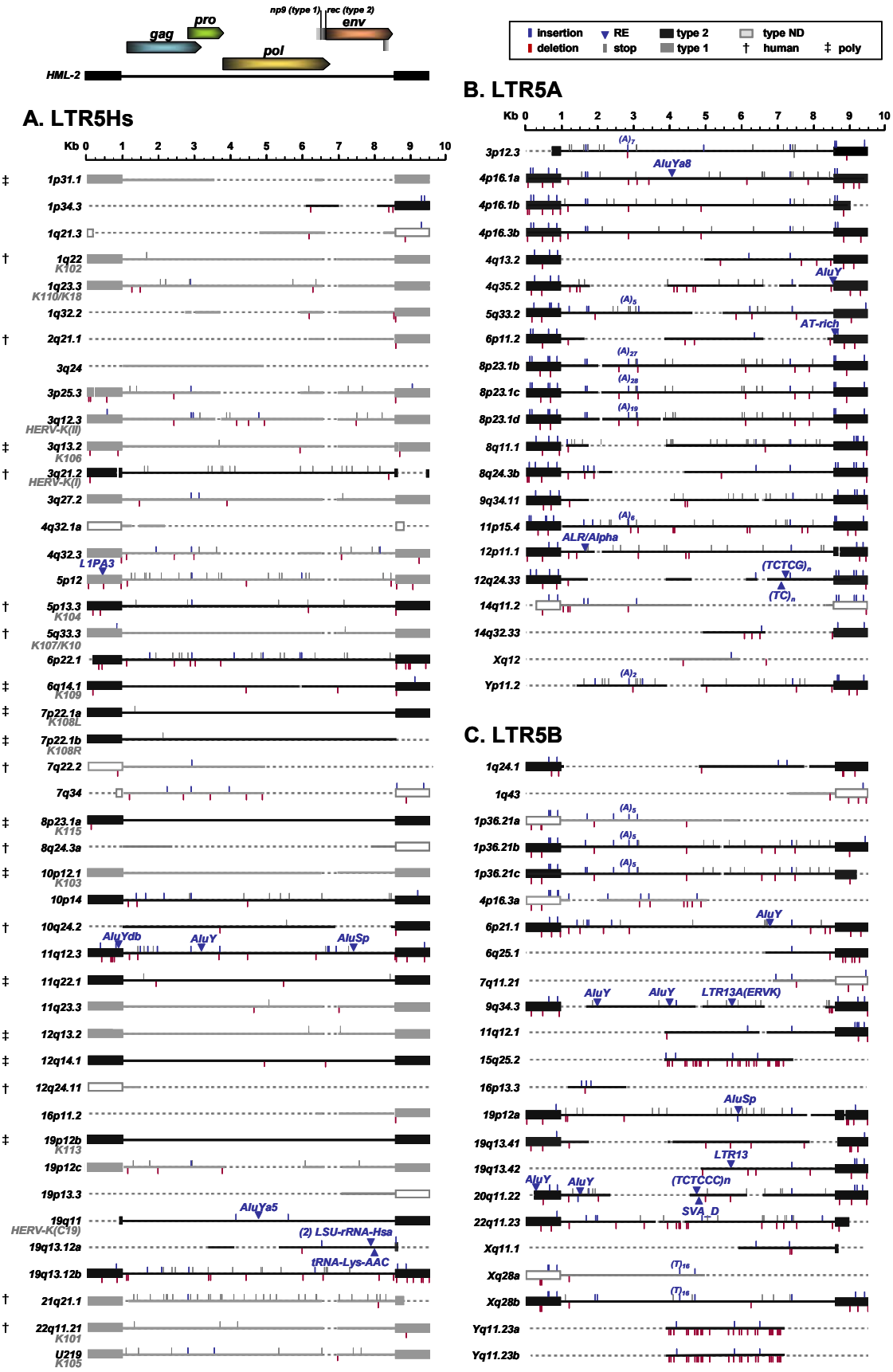

gag
pro
pol
env
np9 (type 1)
rec (type 2)
HML-2
insertion
deletion
RE
stop
type 2
type 1
type ND
human
poly
A. LTR5Hs
B. LTR5A
C. LTR5B
Kb 0 1 2 3 4 5 6 7 8 9 10
1p31.1
1p34.3
1q21.3
1q22 K102
1q23.3 K110/K18
1q32.2
2q21.1
3q24
3p25.3
3q12.3 HERV-K(II)
3q13.2 K106
3q21.2 HERV-K(I)
3q27.2
4q32.1a
4q32.3
L1PA3
5p12
5p13.3 K104
5q33.3 K107/K10
6p22.1
6q14.1 K109
7p22.1a K108L
7p22.1b K108R
7q22.2
7q34
8p23.1a K115
8q24.3a
10p12.1 K103
10p14
10q24.2
AluYdb
AluY
AluSp
11q12.3
11q22.1
11q23.3
12q13.2
12q14.1
12q24.11
16p11.2
19p12b K113
19p12c
19p13.3
AluYa5
19q11 HERV-K(C19)
(2) LSU-rRNA-Hsa
19q13.12a
tRNA-Lys-AAC
19q13.12b
21q21.1
22q11.21 K101
U219 K105
3p12.3
AluYa8
4p16.1a
4p16.1b
4p16.3b
4q13.2
AluY
4q35.2
5q33.2
AT-rich
6p11.2
8p23.1b
8p23.1c
8p23.1d
8q11.1
8q24.3b
9q34.11
11p15.4
ALR/Alpha
12p11.1
(TCTCG)n
12q24.33
(TC)n
14q11.2
14q32.33
Xq12
Yp11.2
1q24.1
1q43
1p36.21a
1p36.21b
1p36.21c
4p16.3a
6p21.1
6q25.1
7q11.21
AluY
AluY
LTR13A(ERVK)
9q34.3
11q12.1
15q25.2
16p13.3
AluSp
19p12a
19q13.41
LTR13
19q13.42
AluY
AluY
(TCTCCC)n
20q11.22
SVA_D
22q11.23
Xq11.1
Xq28a
Xq28b
Yq11.23a
Yq11.23b

Figure 3.3. Schematic representing the inferred structures of the HML-2 proviruses present in humans.

Upper left: Diagram of the HML-2 full-length provirus, indicating the reading frames for the Gag, Pro, Pol, and Env precursor proteins. Also indicated are the protein frames for Rec and Np9, respective to the type 2 and type 1 proviruses. Scale is shown as a reference for individual elements included below.

A-C. Inferred structures of the 89 full-length, near full-length, or partial HML-2 elements identified in this study. Elements are shown according to subgroup, based on the characterization of specific shared nucleotide motifs and phylogeny. The genome location for each element is by chromosomal locus at left, and additional nomenclature is provided when appropriate. Type 2 proviruses are in black, and type 1 in grey. Elements without subtype are white boxed. Insertions/deletions of >3bp are flagged along each element in blue or red, respectively, and stop codons are in grey. Arrowheads represent cases of larger insertions, for example SINEs or LINEs, and referenced. Human-specific elements are indicated by a single crossed bar at far left, and those known to be polymorphic with a double-crossed bar.

Characterization of LTR subgroups within HML-2

In a 2003 study by Buzdin *et al*., [115] the LTRs belonging to the HML-2 group were reported to cluster into specific subgroups based on shared 'diagnostic' features. From the existing genome sequence database at the time, the authors identified 40 highly conserved HML-2 LTRs, which they referred to as "Hs", presumably for 'human specific' or '*Homo sapiens*', as an indicator of the sequence source, or given that 32 of the 40 LTRs were human-specific by PCR [115]. Within the Hs LTRs, the authors noted single nucleotide polymorphisms (SNPs) which effectively divided the Hs LTRs into the "Hs-a" or "Hs-b" subgroups (indicated in Figure 3.4). This classification system was initially implemented into RepBase to algorithmically classify repetitive elements within sequence databases, and then into the RepeatMasker feature within the UCSC Browser. Over time, there has been apparent change to the underlying mechanism(s) for assigning the HML-2 (and likely other types of repetitive sequence) into distinct subgroups, although the exact premise is arguably unclear. Regardless, the most current RepeatMasker program references LTRs belonging to HML-2 as either LTR5-Hs, LTR5A, or LTR5B subgroups.

To better characterize the LTR distribution within the HML-2 group, we attempted to recapitulate the previous classification strategy into our dataset. Proviruses with at least one LTR (81 of 89 total) were aligned and analyzed for the presence of each subgroup-specific feature, as described below. A total of 8 elements without an associated LTR were excluded here, but have been included in other analyses when appropriate, given the presence of >50% internal sequence (also refer to Figure 3.3). For

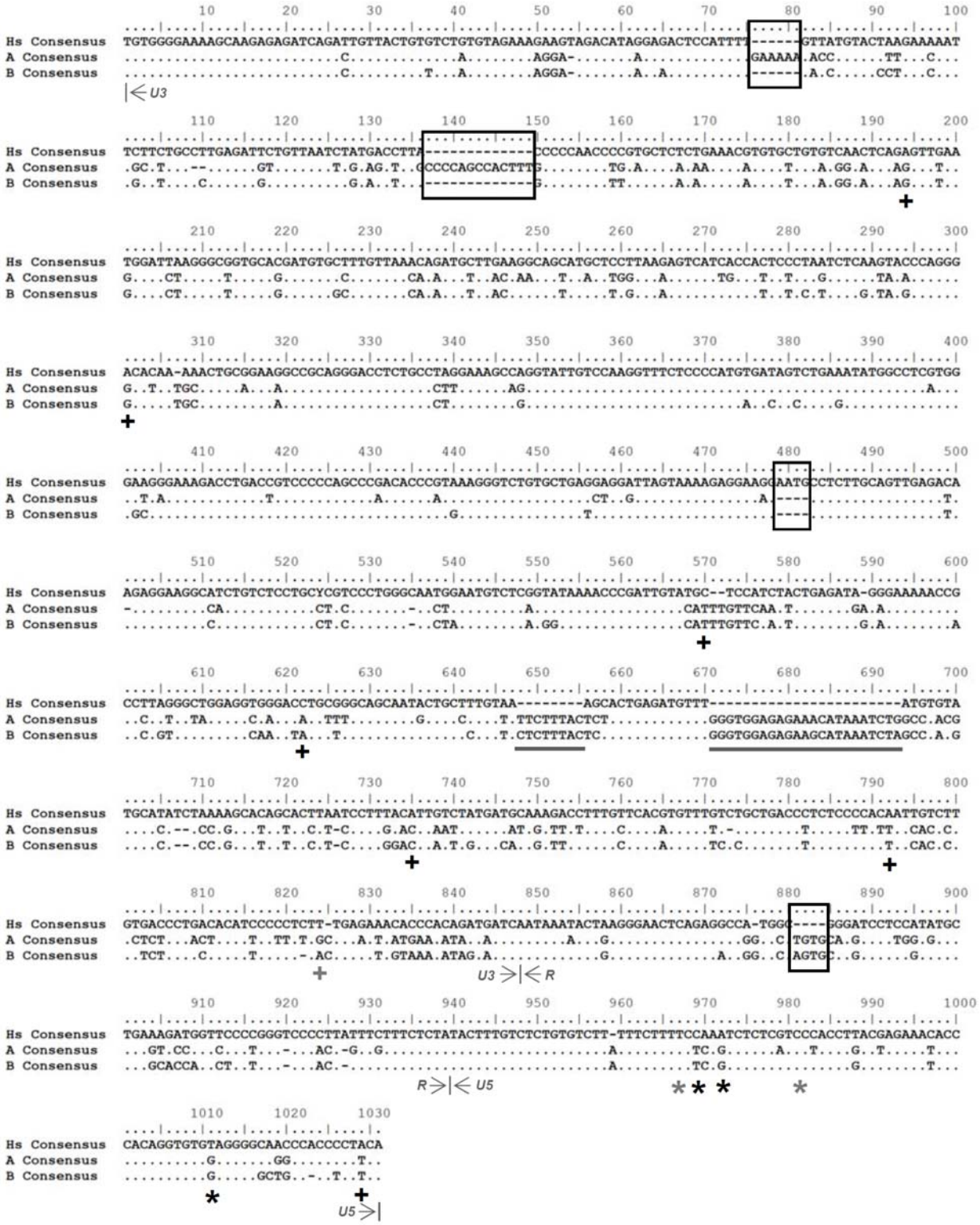

Hs Consensus
A Consensus
B Consensus
U3
R
U5

Figure 3.4. Identification of subgroup-specific nucleotide motifs within the LTRs of HML-2 proviruses.

Shown is an alignment of the consensus sequences derived from the LTR5-Hs, LTR5A, and LTR5B subgroup proviruses. The 5' and 3' LTR from each the provirus were aligned in ClustalW and adjusted manually using in BioEdit v.7.0.9.0. Elements for which single and both LTRs were included. The subgroup consensus LTR sequences were generated with reference to the most common base at each position and manually aligned in BioEdit. The U3, R, and U5 boundaries are labeled as such. Crosses and asterisks were used to denote the positions of SNPs previously reported as specific to each subgroup [115]: Crosses indicate the putative subgroup A-specific SNPs, and asterisks indicate those reported as specific to the LTR5-Hs. Those SNPs found to be consistent between the prior study and this analysis are in black; SNPs not upheld as specific to the LTR5-Hs from our results are in grey.

simplicity, and to stay consistent with the most recent human genome builds, the current nomenclature was retained for the analyses here.

From the total full-length alignment of HML-2, individual proviruses were referenced according to RepeatMasker, and tentatively divided into the LTR5-Hs, A, and B groups. Nucleotide sequences for the 5' and 3' LTRs were extracted and manually aligned, from which the consensus was derived per subgroup. We then performed two searches to verify the subgroup features reported by Buzdin *et al*., the first within the total aligned LTRs, and another based on the consensus sequences derived from each subgroup (Figure 3.4). Within the larger set of total individually aligned LTRs, SNPs designated by Buzdin *et al*. were generally accurate, however the majority of sites were polymorphic and represented by multiple bases. Most SNPs appeared to discriminate the LTR5-Hs from A and B, however the reverse was not observed, and no SNPs were found to distinguish B or A from Hs (not shown). Inspection of the consensus subgroup sequences provided clearer support of this finding, verifying 7 of 8 SNPs from the 'Hs', and 3 of 5 SNPs from 'Hs-b', though again without discrimination of either LTR5A or B from the Hs elements (indicated and detailed in Figure 3.4).

The differences observed in our dataset from previous reports were likely due to the present incorporation of a larger sample size [94, 115], and the consequent analysis of many older and more divergent members of the HML-2 group. To better characterize the LTR subgroups within the whole of HML-2, we searched for and identified alternate sequence motifs restricted to each group (specific sites are provided in Figure 3.4). Denoting the LTR5-Hs group was a 4bp insertion (C/AATG) identified in 37 of 43 proviruses (63 of 69 LTRs) at the consensus base position 479-483. The LTR5A

contained a subgroup-specific insertion ($GC_4AGCCACT_3$) from consensus bases 150 to 163 present in all elements, and a second insertion (GA_5) in 14 of 20 proviruses (24 of 33 LTRs). Features specific to the LTR5B were not located, however sequences shared by most members of LTR5A and B, but absent in the -Hs, were observed, for example the 4bp insertion (A/TTGT) at consensus position 881. Also were two larger insertions from positions 648 and 671 that were specific to LTR5A and B, although the specific boundaries of each site were difficult to determine due to high sequence divergence within this region (underlined in Figure 3.4).

Direct comparison of the LTR subgroup assignments for each provirus using our classification system, in combination with the system implemented by RepeatMasker [115], were largely in agreement, but not without exception. Specifically, the LTRs from the 8q11.1 and 4q13.2 proviruses are assigned to different subgroups by RepeatMasker: 8q11.1 as having a 5'LTR5B and 3' LTR5A, while the opposite was seen for 4q13.2 (5'LTR5A and 3'LTR5B). However, the LTRs from both elements are consistent with LTR5A according to our classification. The possibility of a recombination between these elements, generating the full-length A/B allele, was ruled out by comparing the direct repeats flanking both 8q11.1 and 4q13.2, which were found to be intact and identical, making this scenario unlikely (data not shown). Furthermore, phylogenetic analysis of these elements demonstrated the 5' and 3' LTRs from each provirus grouped together, as would be expected for the LTRs of a single provirus (below; asterisked in Figure 3.5). These observations provide support that both proviruses were derived from the LTR5A subgroup, with the added implication that each was algorithmically mis–annotated within the UCSC database.

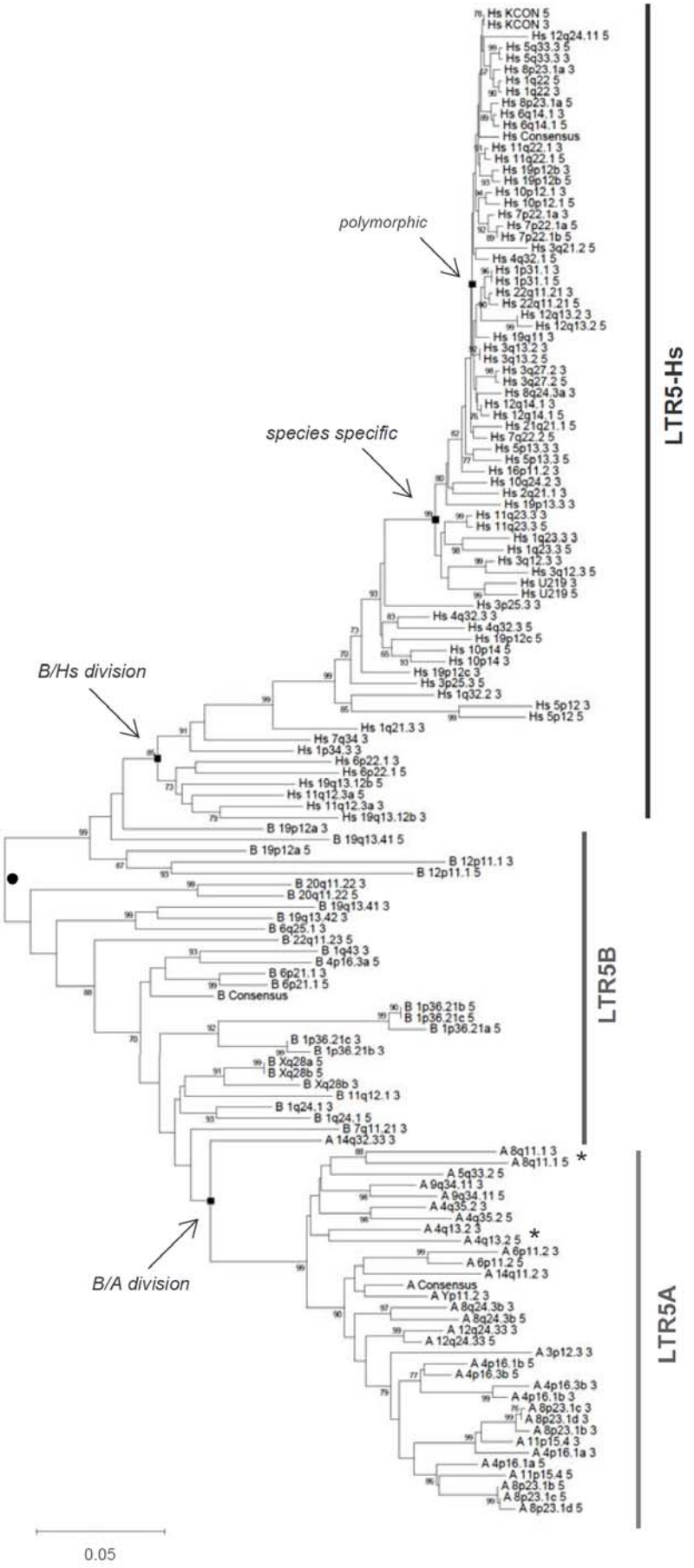

Hs KCON 5
Hs KCON 3
Hs 12q24.11 5
Hs 5q33.3 5
Hs 5q33.3 3
Hs 8p23.1a 3
Hs 1q22 5
Hs 1q22 3
Hs 8p23.1a 5
Hs 6q14.1 3
Hs 6q14.1 5
Hs Consensus
Hs 11q22.1 3
Hs 11q22.1 5
Hs 19p12b 3
Hs 19p12b 5
Hs 10p12.1 3
Hs 10p12.1 5
Hs 7p22.1a 3
Hs 7p22.1a 5
Hs 7p22.1b 5
Hs 3q21.2 5
Hs 4q32.1 5
Hs 1p31.1 3
Hs 1p31.1 5
Hs 22q11.21 3
Hs 22q11.21 5
Hs 12q13.2 3
Hs 12q13.2 5
Hs 19q11 3
Hs 3q13.2 3
Hs 3q13.2 5
Hs 3q27.2 3
Hs 3q27.2 5
Hs 8q24.3a 3
Hs 12q14.1 3
Hs 12q14.1 5
Hs 21q21.1 5
Hs 7q22.2 5
Hs 5p13.3 3
Hs 5p13.3 5
Hs 16p11.2 3
Hs 10q24.2 3
Hs 2q21.1 3
Hs 19p13.3 3
Hs 11q23.3 3
Hs 11q23.3 5
Hs 1q23.3 3
Hs 1q23.3 5
Hs 3q12.3 3
Hs 3q12.3 5
Hs U219 3
Hs U219 5
Hs 3p25.3 3
Hs 4q32.3 3
Hs 4q32.3 5
Hs 19p12c 5
Hs 10p14 5
Hs 10p14 3
Hs 19p12c 3
Hs 3p25.3 5
Hs 1q32.2 3
Hs 5p12 3
Hs 5p12 5
Hs 1q21.3 3
Hs 7q34 3
Hs 1p34.3 3
Hs 6p22.1 3
Hs 6p22.1 5
Hs 19q13.12b 5
Hs 11q12.3a 5
Hs 11q12.3a 3
Hs 19q13.12b 3
B 19p12a 3
B 19q13.41 5
B 19p12a 5
B 12p11.1 3
B 12p11.1 5
B 20q11.22 3
B 20q11.22 5
B 19q13.41 3
B 19q13.42 3
B 6q25.1 3
B 22q11.23 5
B 1q43 3
B 4p16.3a 5
B 6p21.1 3
B 6p21.1 5
B Consensus
B 1p36.21b 5
B 1p36.21c 5
B 1p36.21a 5
B 1p36.21c 3
B 1p36.21b 3
B Xq28a 5
B Xq28b 5
B Xq28b 3
B 11q12.1 3
B 1q24.1 3
B 1q24.1 5
B 7q11.21 3
A 14q32.33 3
A 8q11.1 3
A 8q11.1 5
A 5q33.2 5
A 9q34.11 3
A 9q34.11 5
A 4q35.2 3
A 4q35.2 5
A 4q13.2 3
A 4q13.2 5
A 6p11.2 3
A 6p11.2 5
A 14q11.2 3
A Consensus
A Yp11.2 3
A 8q24.3b 3
A 8q24.3b 5
A 12q24.33 3
A 12q24.33 5
A 3p12.3 3
A 4p16.1b 5
A 4p16.3b 5
A 4p16.3b 3
A 4p16.1b 3
A 8p23.1c 3
A 8p23.1d 3
A 8p23.1b 3
A 11p15.4 3
A 4p16.1a 3
A 4p16.1a 5
A 11p15.4 5
A 8p23.1b 5
A 8p23.1c 5
A 8p23.1d 5
polymorphic
species specific
B/Hs division
B/A division
LTR5-Hs
LTR5B
LTR5A
*
*
0.05

Figure 3.5. Neighbor-joining tree constructed with the nucleotide sequences of LTRs from full-length and near full-length HML-2 proviruses.

The 5' and 3' LTRs were included if observed as associated with proviruses; single LTRs were also included. The full-length alignment was constructed with ClustalW and manually aligned in BioEdit v.7.0.9.0. A neighbor-joining tree was constructed and edited using MEGA v.4.0 with the Kimura 2-parameter (K2P) distance correction. Bootstrapping was performed from 5,000 replicates and randomly seeded. Values above 65 are shown at the respective nodes. Superimposition of subgroup information onto the phylogeny is indicated at right by bars dividing the LTR5-Hs, LTR5A, or LTR5B elements. The nodes for each respective subgroup within the tree are indicated with arrows and labeled at left. Attempts to align *Beta*-like HERVs or exogenous *Betaretroviruses* related to the HML-2 to provide an outgroup were hampered by their low sequence similarity, and as a result the tree was rooted on its midpoint (filled black circle).

With reference to the subgroup-specific sequence motifs as detailed above, 43 proviruses were classified to the LTR5-Hs subgroup (~53%), 20 were assigned to the LTR5A subgroup (~25%), and 18 to the LTR5B elements (~22%). The observed proportions differed from previous estimates, in which the LTR5-Hs constituted as many as 63% of HML-2 proviruses, but with the caveat on the analysis of relatively few loci (14 in all) from early genome builds [115].

LTR-based phylogenetic analysis

Due to the mechanism of reverse transcription, the LTRs of a provirus are identical at the time of integration, and subsequently evolve independently within the host. Because of this property, the LTRs belonging to a single provirus will be more similar to one another than to any other element, and so should always group together within a tree. Deviation from this predicted branching pattern indicates recombination involving those proviruses, for example from recombination, gene conversion, or duplication [63-65, 71]. Consequently, the topology and branch lengths within an LTR-based phylogeny reflect both the exogenous and endogenous evolution of a group of proviruses. The internal branches infer evolution of the virus during exogenous replication, prior to germline integration, and the terminal branches reflect the evolution of a provirus that has occurred following its stable integration into the host genome. In order to further examine the overall branching patterns and evolution of proviruses within the HML-2 group, we performed a phylogenetic analysis of the proviral-associated LTRs (Figures 3.5 and 3.6).

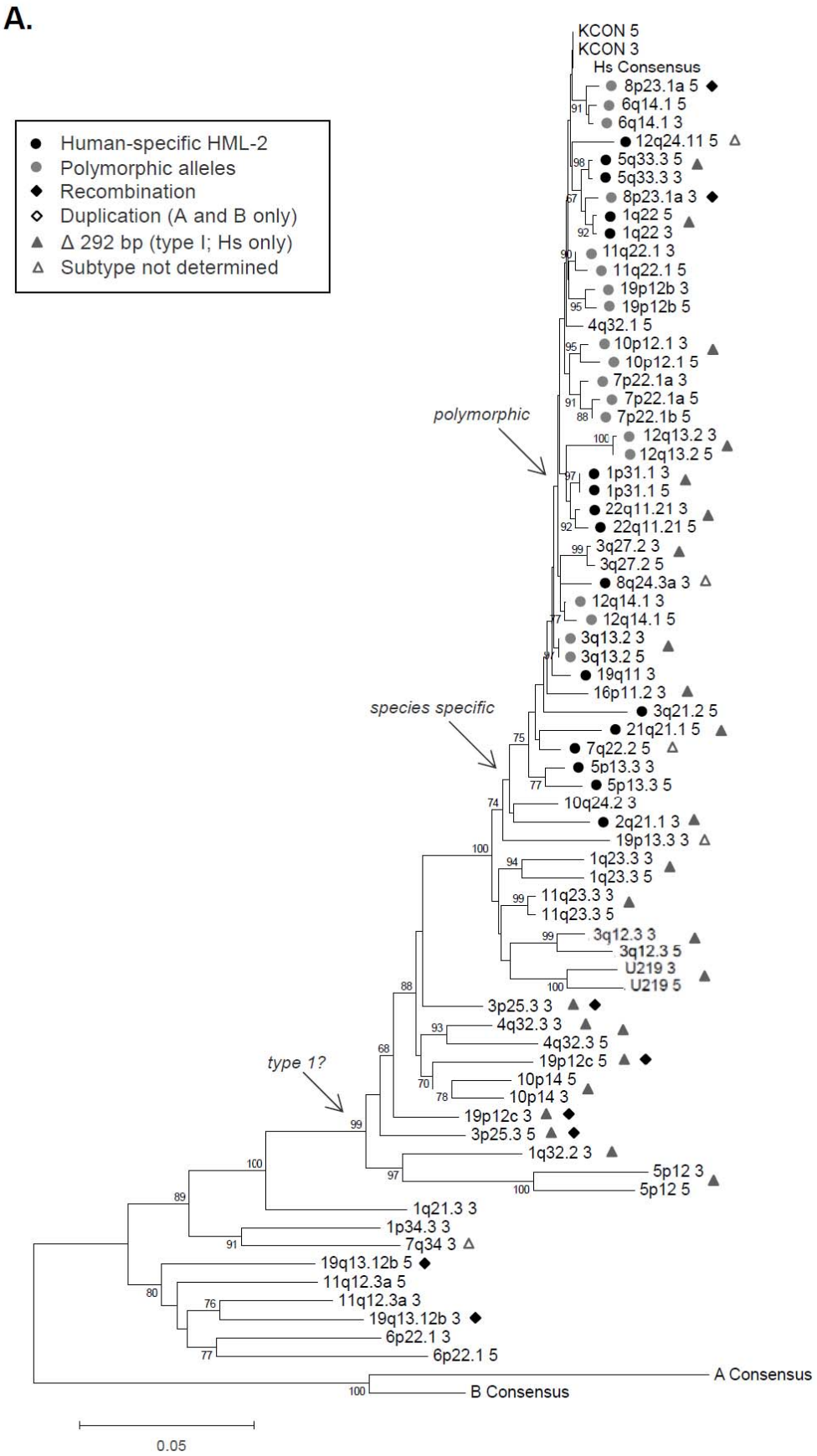
Phylogenetic support for HML-2 subgroup distribution

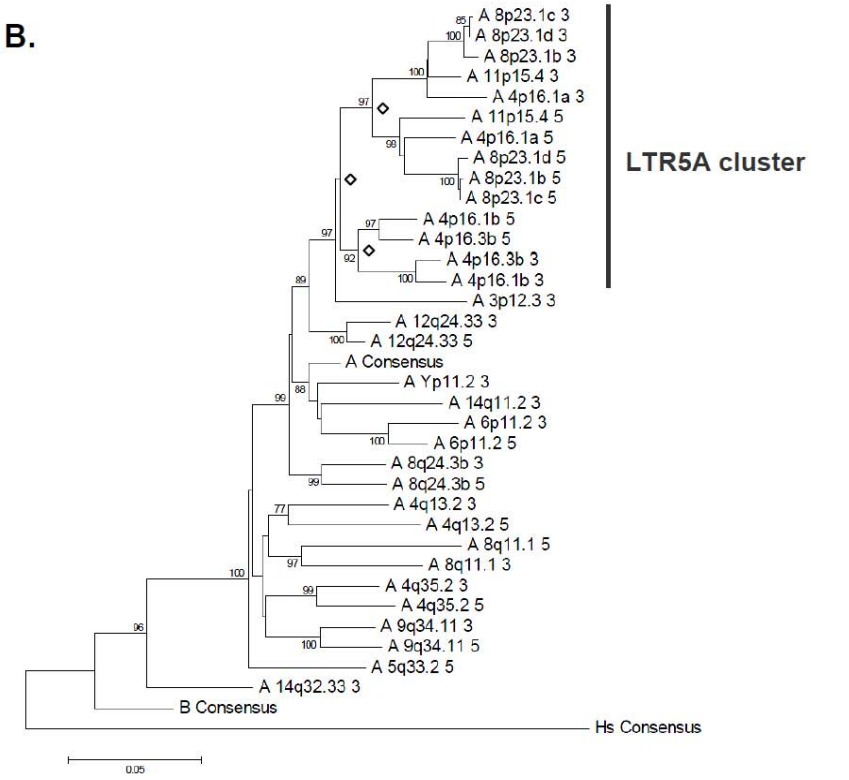
Initially, a single neighbor-joining tree was generated from the alignment of total LTRs from proviruses with either one or both LTRs present (Figure 3.5). This tree was used as a reference to first analyze the overall branching patterns within HML-2 as a group, and to provide support for the separation of each LTR subgroup. To give better direction to the tree, we attempted to include an outgroup from other HERVs closely related to HML-2, including the canonical HML-1 to -10, and the exogenous MMTV and

JSRV. However, we were prevented from identifying a reliable outgroup due to their low similarity, and the tree was instead rooted at its midpoint. The LTRs were then labeled according to their respective subgroup of LTR5-Hs, A, or B, as determined above. Subsequently, individual trees were generated for the LTR5-Hs, A, and B proviruses and analyzed with reference to the consensus sequences of the remaining subgroups (Figure 3.6).

The primary tree is shown in Figure 3.5, in which two major lineages were observed. The largest (upper) clade contained the most recent integrations, including all human-specific members, and those with polymorphic alleles, nearer the tips of the tree with the shortest branch lengths (indicated respectively by black and grey triangles). Within the second lineage (lower) were most of the evolutionarily older HML-2 proviruses, many of which are shared among most primates, with little evidence of activity in humans following the *Homo-Pan* divergence (refer to Figure 3.3, and also R. Subramanian, personal communication) [64, 94, 115]. Overall, the observed phylogeny was consistent with other reports in the arrangement of previously described proviruses [64, 71, 94, 115]. However the addition of LTRs from 20 previously undescribed proviruses markedly enhanced the tree topology, thus providing a more informative representation of the evolutionary dynamics within the HML-2 group.

We searched for evidence of inter-element recombination, as specific properties of shared LTRs allow for inference of genomic recombination and/or rearrangements through their phylogenetic analysis [64, 65, 71]. The 5' and 3' LTRs of a provirus are identical at integration, and as such will group as sister taxa in a phylogenetic tree. Thus, non-paired 5' and 3' LTRs of an individual provirus can indicate


A.
Human-specific HML-2
Polymorphic alleles
Recombination
Duplication (A and B only)
Δ 292 bp (type I; Hs only)
Subtype not determined
KCON 5
KCON 3
Hs Consensus
8p23.1a 5
6q14.1 5
6q14.1 3
12q24.11 5
5q33.3 5
5q33.3 3
8p23.1a 3
1q22 5
1q22 3
11q22.1 3
11q22.1 5
19p12b 3
19p12b 5
4q32.1 5
10p12.1 3
10p12.1 5
7p22.1a 3
7p22.1a 5
7p22.1b 5
12q13.2 3
12q13.2 5
1p31.1 3
1p31.1 5
22q11.21 3
22q11.21 5
3q27.2 3
3q27.2 5
8q24.3a 3
12q14.1 3
12q14.1 5
3q13.2 3
3q13.2 5
19q11 3
16p11.2 3
3q21.2 5
21q21.1 5
7q22.2 5
5p13.3 3
5p13.3 5
10q24.2 3
2q21.1 3
19p13.3 3
1q23.3 3
1q23.3 5
11q23.3 3
11q23.3 5
3q12.3 3
3q12.3 5
U219 3
U219 5
3p25.3 3
4q32.3 3
4q32.3 5
19p12c 5
10p14 5
10p14 3
19p12c 3
3p25.3 5
1q32.2 3
5p12 3
5p12 5
1q21.3 3
1p34.3 3
7q34 3
19q13.12b 5
11q12.3a 5
11q12.3a 3
19q13.12b 3
6p22.1 3
6p22.1 5
A Consensus
B Consensus
0.05
polymorphic
species specific
type 1?


B.
A 8p23.1c 3
A 8p23.1d 3
A 8p23.1b 3
A 11p15.4 3
A 4p16.1a 3
A 11p15.4 5
A 4p16.1a 5
A 8p23.1d 5
A 8p23.1b 5
A 8p23.1c 5
A 4p16.1b 5
A 4p16.3b 5
A 4p16.3b 3
A 4p16.1b 3
A 3p12.3 3
A 12q24.33 3
A 12q24.33 5
A Consensus
A Yp11.2 3
A 14q11.2 3
A 6p11.2 3
A 6p11.2 5
A 8q24.3b 3
A 8q24.3b 5
A 4q13.2 3
A 4q13.2 5
A 8q11.1 5
A 8q11.1 3
A 4q35.2 3
A 4q35.2 5
A 9q34.11 3
A 9q34.11 5
A 5q33.2 5
A 14q32.33 3
B Consensus
Hs Consensus
0.05
LTR5A cluster

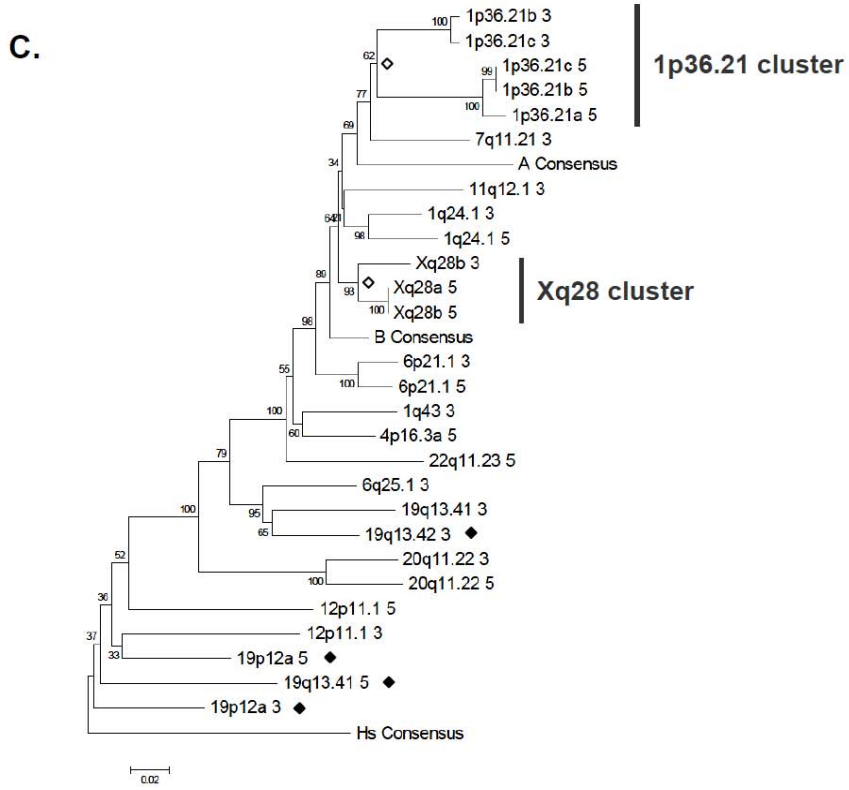

C.
1p36.21b 3
1p36.21c 3
1p36.21c 5
1p36.21b 5
1p36.21a 5
7q11.21 3
A Consensus
11q12.1 3
1q24.1 3
1q24.1 5
Xq28b 3
Xq28a 5
Xq28b 5
B Consensus
6p21.1 3
6p21.1 5
1q43 3
4p16.3a 5
22q11.23 5
6q25.1 3
19q13.41 3
19q13.42 3
20q11.22 3
20q11.22 5
12p11.1 5
12p11.1 3
19p12a 5
19q13.41 5
19p12a 3
Hs Consensus
0.02
1p36.21 cluster
Xq28 cluster

Figure 3.6. Neighbor-joining trees representing the LTR5-Hs, LTR5A, and LTR5B subgroups of HML-2 proviruses.

The neighbor-joining trees were generated using aligned LTR sequences divided by subgroup from proviruses with either one or both LTRs present. Each tree was constructed and edited using MEGA v.4.0 with the Kimura 2-parameter (K2P) distance correction. Bootstrapping was performed from 5,000 replicates and randomly seeded. Values above 65 are shown at the respective nodes. Consensus sequences of the remaining subgroups were included for use of outgroup and support in topology.

A. The LTR5-Hs tree was rooted with the branch leading to the consensus LTR5A and B subgroups. Nodes are indicated for the clades representative of the type 1, species-specific (detected in humans only), and polymorphic elements. Filled circles at branch termini denote human-specific proviruses; grey are polymorphic and black detected in other species. Indicated at right by grey triangles are the type 1, and open triangles designate non-typed elements. Proviruses with evidence of recombination are denoted with filled black diamonds.

B. The LTR5A tree was rooted with the LTR5-Hs consensus. Duplicates are indicated at right and by open diamonds at nodes; only the first duplications events leading to either inner lineage are shown.

C. The LTR5B tree was also rooted with the LTR5-Hs consensus. Duplications are indicated as for (B) and elements with evidence of recombination by filled diamonds.

recombination between distinct elements (Figure 3.7). There were 6 examples of non-paired LTRs, from proviruses located at chromosomal positions 3p25.3, 8p23.1a (K115), 11q12.3, 19p12a, 19p12c and 19q13.12b (indicated with filled diamonds in Figure 3.6 A). We confirmed their homology to previously described elements [64, 95] by BLAT of each to the earliest available genome build (July 2003, NCBI34/hg16). With particular attention to the newly identified elements still associated with both the 5' and 3' LTRs, we observed clustering typical of paired LTRs that have evolved in the absence of rearrangement of recombination, or without sufficient gene conversion to alter the overall sequence similarities between the pairs. Thus, among the newly identified elements, none was observed with unexpected clustering of the LTRs, suggesting most recombination and/or gene conversion events are observed within the LTR5-Hs, but are less frequent in the context of the HML-2 elements as a group.

Superimposition of subgroup information for each provirus added support to our strategy for HML-2 classification. The clustering of each subgroup within the LTR-based tree indicated the largest clade was represented by the LTR5-Hs subgroup, and the secondary lineage divided into the LTR5A and B subgroups. Also within the tree, each provirus clustered most closely with other members of the same subgroup, resulting in the separation between the three subgroups at distinct nodes (each is arrowed in Figure 3.5). The LTR5-Hs also contained the distinct node giving rise to the human-specific elements, and again to those with polymorphic insertions. Overall, the tree topology and clustering of subgroups suggests exogenous viruses unique to each subgroup contributed to germline integrations leading to the observed HML-2 phylogeny.

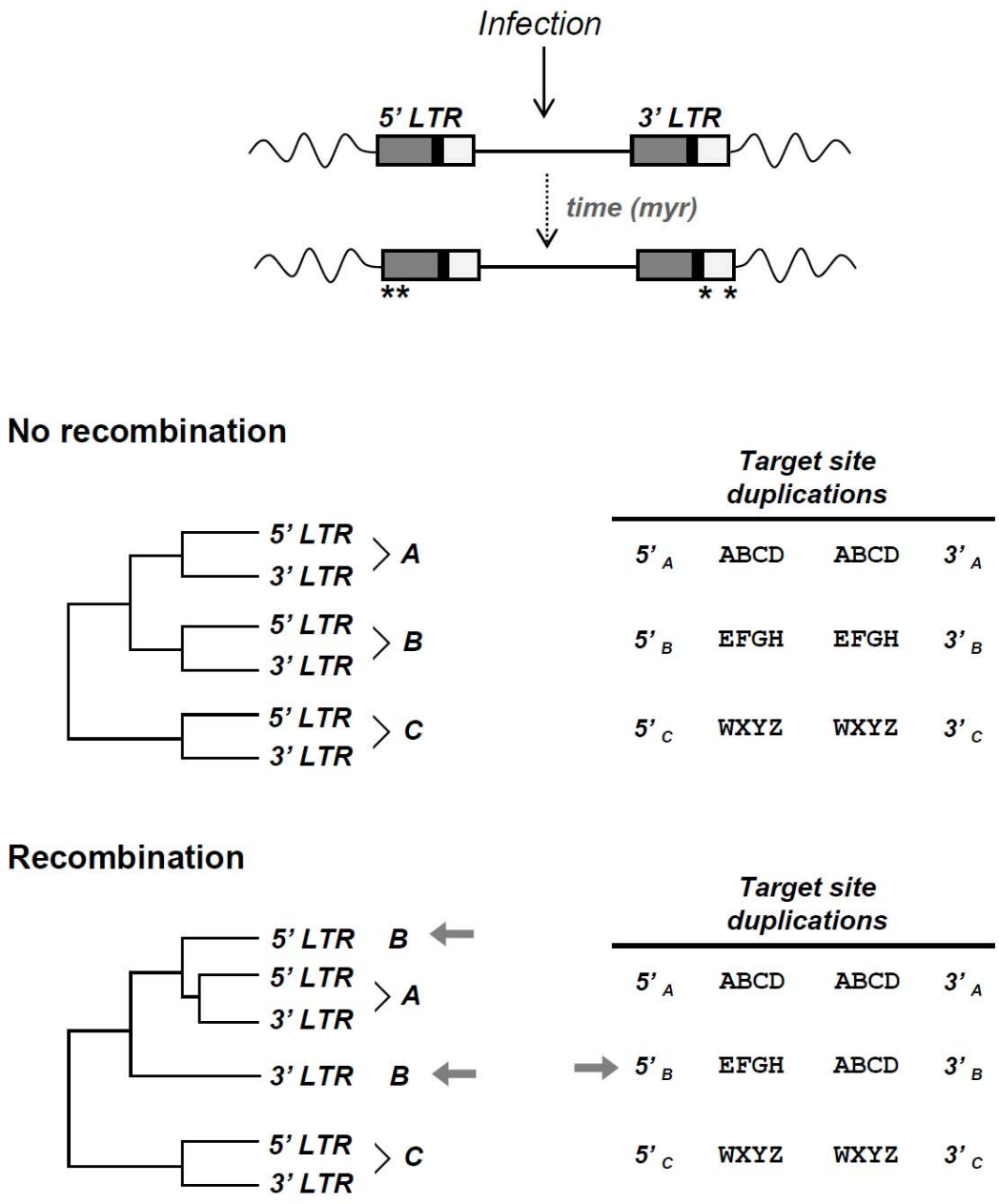


Figure 3.7. Representation of LTR evolution and the detection of recombination. An endogenous provirus is shown following infection of the germline. Mutations are indicated by asterisks below wither LTR over time (in millions of years). LTR trees indicate pairing of the 5' and 3' LTRs of three given elements over time in the absence (upper tree) or presence (lower tree) of recombination. The representative target site duplications are indicated at right, with gray arrows indicating the properties detected following recombination.

From the total HML-2 proviruses, 8 were apparently without LTR association, and we tentatively categorized each to a subgroup based on the analysis of their internal reading frames. For this purpose, we compared a region of the *pol* gene present in most HML-2 (65 in all) in a phylogeny with reference to representative members of HERV-K (HML-1 through HML-10) and the exogenous *Betaretrovirus* groups MMTV and JSRV. For the HML-2 with *pol* deletions, the analysis was repeated from alignments of the first ~1.8 kb of *gag,* and ~1 kb of *env* corresponding to the SU region. Briefly, each tree displayed the same overall branching pattern and evolutionary relationship as the LTR-based tree, with supported clustering of each subgroup (data not shown). The agreement between trees, and added inference from sequence similarities and shared structures within each subgroup (Figure 3.3) added confidence in subgroup assignment for the 8 HML-2 with LTRs no longer intact. In all, two elements were added to the LTR5-Hs, one to the LTR5A, and 5 to LTR5B, for a total distribution of 45 proviruses belonging to LTR5-Hs (~51%), 20 LTR5A (~22%), and 24 LTR5B (~27%).

Evolution and relative times of formation of the HML-2 subgroups

The primary HML-2 phylogeny was of interest in that it suggested the LTR5B proviruses were ancestral to both the Hs and A groups. Prior analyses of the HML-2 group have had primary interest in human-specific elements, and in most cases have included LTR5A or B proviruses for the purposes of providing an outgroup (but also see Belshaw, *et al*. 2004). In doing so, the relationship between the subgroups has been unclear with respect to the LTR5A and B elements. Age estimates and phylogenetic inference for individual proviruses have supported the LTR5-Hs as the evolutionarily younger group, while the A and B groups are the result of less recent germline integrations [94, 115]. That the LTR5B was ancestral to both -Hs and A has not been previously suggested.

To further investigate the branching pattern we observed, and to support the uniqueness of each subgroup, individual phylogenies were generated (LTR5-Hs, A, or B) and rooted with reference to the most distant subgroup consensus (Figure 3.6 A, B, and C). Each subgroup tree was well-supported and displayed congruency with the primary tree. While each subgroup tree was generated with the same sequence information, and as a result demonstrated the same branching and topology, the presentation of each subgroup individually allows for a clearer interpretation of the evolutionary relationships within each group in the context of the consensus sequence of the remaining HML-2 proviruses.

Consistent with a number of previous reports, the LTR5-Hs proviruses were observed in a monophyletic clade in a tree rooted with the branch leading to the consensus for the LTR5A and B (Figure 3.6 A). Within this tree, each provirus has been

labeled as human-specific (inner black circles) or as polymorphic (grey circles). All human-specific proviruses were located within a well-supported lineage (arrowed). The majority of elements within this clade had relatively short branch lengths, especially in comparison with the representative phylogenies of the LTR5A and B subgroups, providing support for their recent germline integration. Also within this lineage, just 5 proviruses have been demonstrated to share orthologous insertions in other primates, as from previous work in our lab and others [64, 65, 94]. Also, the most distant ape species in which each has been detected is the gorilla, again indicating the relatively recent activity of this particular clade in primates.

The branch representing the LTR5A and B consensus sequences indicated a closer relationship of the LTR5-Hs with the B subgroup, consistent with the prediction that the LTR5-Hs and A evolved separately from the B subgroup. Also in agreement was the placement of the LTR5-Hs and B consensus sequences in LTR5A phylogeny (Figure 3.6 B), in which the B subgroup was less divergent than the Hs. Finally, the tree representing the LTR5B appeared ancestral to both the LTR5A and B consensus sequences, with the –Hs exhibiting the most change (Figure 3.6 C).

With the total number HML-2 proviruses having been under represented in earlier studies, age estimations per subgroup were placed fairly recently, for example one analysis in 2001 calculated an average time of formation of ~5.8 and ~10.3 mya for LTR5A and B, respectively [115]. Indeed, it has since been confirmed that the majority of LTR5A and B proviruses have shared loci among primates, with some even detected in Old World monkeys and no described human-specific elements of the LTR5A or LTR5B [64, 65]. By virtue of the mentioned properties of the LTRs proviruses at

integration, we estimated the ages for 51 elements with intact LTRs that appeared to have evolved in the absence of rearrangement. Those reported to have been involved in ecoptic rearrangements were excluded from the analysis [64, 65]. Using the method previously described by Lebedev, *et al*. [189], the total number of nucleotide differences between LTRs were treated as a percent of the LTR length, with insertions and deletions counted as single changes, and the resulting frequency normalized to the mutation rate of 2x(0.13%) per mya per provirus. The value of 0.13% changes per mya corresponds to the averaged mutation rates previously estimated from the intrabranch divergences of multiple HERVs groups, and takes into consideration the average times of species' divergences during primate evolution [73]. Also, this averaged rate is reasonably in agreement to the estimate of 0.12% for the HERV-K (HML-2), following the first characterizations of the group [67]. The two rates are in reasonable agreement, and in general, values that correspond to, or are close to 0.13% changes per mya have been used for these purposes [73].

Subsequent age estimations are provided in Table 3.1 for individual elements, including the number of base changes between the LTRs from each analyzed provirus as a reference. With an estimated divergence rate of 0.13% changes per mya, a single LTR is predicted to accumulate a base change every ~0.8 million years (myr) [189]. Therefore, for elements with very few changes between LTRs, the age estimations according to this method can be less informative, given the prediction above for the accumulated base changes per myr. For LTRs that have no observed differences between LTRs, the value of 0.8 myr becomes the lower limit for age estimation. To avoid this issue we used a cut-off of <2 mya for our age estimations. Overall, the values obtained for the times of

integration ranged from less than 2 mya to greater than 40 mya, depending on the provirus, and were consistent with previous reports [89, 94, 186]. Comparison of the individual elements' ages and subgroups gave a general indication that the older HML-2 were more common to the LTR5A and B, with the opposite for the LTR5-Hs. For general support of this trend, we plotted the times of integrations as a function of subgroup, shown as a boxplot in Figure 3.8.

As mentioned, the LTR5B elements appear to have had the least recent times of integration, at an average time of formation at 14.7 mya, and prior to the activities of the LTR5-Hs and A elements. Supporting their more recent activity, the LTR5A and –Hs subgroups were estimated to have times of formation averaging around 8.4 and ~5.1 mya, respectively. The inferred times of integration of these two subgroups are also consistent with their positioning within the LTR-based phylogeny, as well as in their apparent 'overlap' in evolutionary activity following emergence from the LTR5B ancestors. Interestingly, the LTR5A and B subgroups appear to have had most of the proviral formation during a more narrow time span, particularly in comparison to the collective LTR5-Hs elements. If we consider the species distributions of elements within each of the subgroups [64, 65], from those sites that have been tested, the LTR5-Hs proviruses appear to have a broader host range, and have been detected in the gibbon (19q13.41b), but not the macaque; only LTR5B elements have been detected in Old World monkeys (the 20q11.23 and 6p21.1 elements). The most recently diverged species for the LTR5B subgroup was from the orangutan (1q24.1), and for the LTR5A, the gorilla (12q24.33) [64, 65]. A possible interpretation of these observations is that, since their appearance the

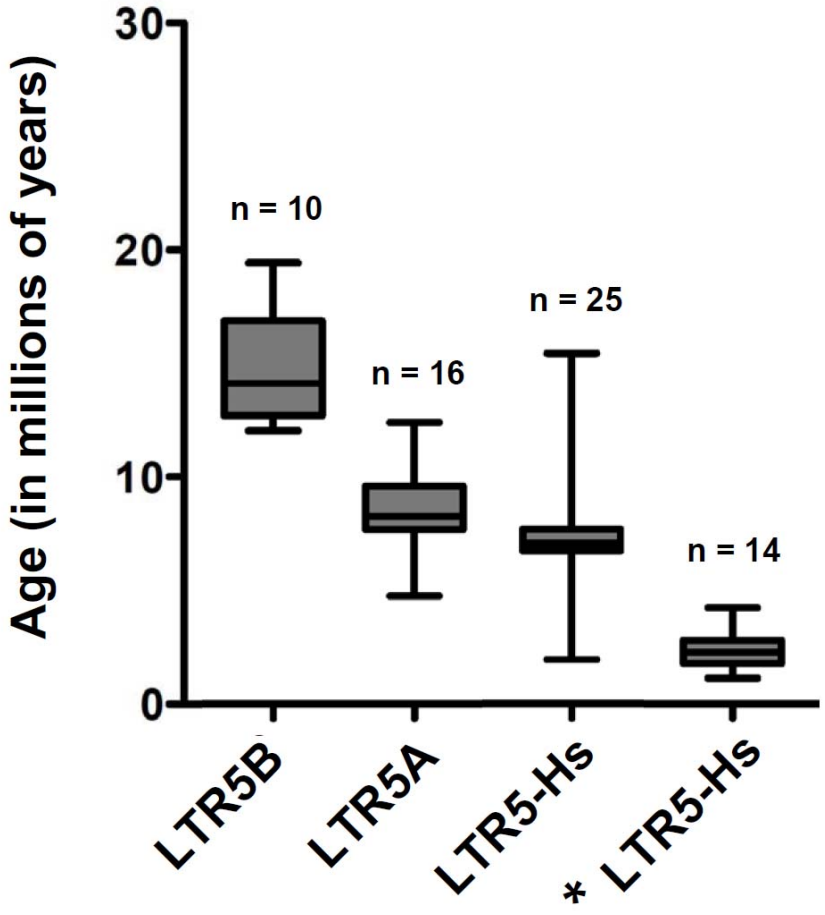


Figure 3.8. Box-plot representation of the age estimates for HML-2 subgroups by sequence comparison of proviral LTRs.
Nucleotide differences between the LTRs of individual proviruses were used to infer evolutionary change following germline insertion for each element. Calculations were performed as described by Lebedev, *et al* [189]. The total base changes between the 5' and 3' LTRs for each provirus were scored as a percent of the LTR length, and normalized to the substitution frequency of 2x0.13%. Only proviruses with both LTRs were included; calculations were not performed for elements with evidence of recombination, as inferred from the aberrant clustering of LTRs. Boxplots were generated for the estimated ages of the LTR5-Hs, A, and B subgroups; also included is the plot of elements detected only in humans; the asterisk indicates elements specific to humans. Adapted from R.P.S., J.H.W., *et al*., in preparation.

LTR5-Hs have retained activity through evolution of the most recent primates, with the independent extinctions of the LTR5B and A subgroups, in that order.

Together with the phylogenetic analyses per subgroup, these results provide evidence that the LTR5-Hs and A clades arose independently from the LTR5B subgroup, and that each was active in the lineage leading to humans. The apparent skew in subgroup distribution toward the LTR5-Hs elements indicates the relative success of the subgroup in recent primate evolution, particularly in humans, whereas the LTR5A and LTR5B have had comparatively little germline activities. Indeed, exclusion of the 27 human-specific elements leaves just 18 proviruses (~20%) from the LTR5-Hs group, comparable to the total number of insertions representing the LTR5A and B subgroups. Furthermore, at least three clusters within the LTR5A and B subgroups exhibit aberrant clustering indicative of an increase in copy number via duplication, raising the possibility that the copy number within these groups has not been from exogenous germline activity alone.

Analysis of the HML-2 proviruses by subtype

Given the representative proportions of the LTR subgroups within HML-2, we asked whether the distribution of type 1 or type 2 might also be skewed to a particular subgroup. As described above, the HML-2 proviruses were previously shown to fall into one of two subtypes, discriminated by the presence or absence of a 292 bp deletion at the *pol-env* junction (Figures 3.3 and 1.14) [190]. For historical reasons, the subtypes are referred to as either type 2 or 1, respectively. It has been well-established that type 2 proviruses encode (or at one time encoded) a spliced reading frame for the accessory protein Rec, a protein involved in the transport of unspliced mRNAs from the nucleus to

the cytoplasm [123, 191]. The *rec* splice site is located within the Δ292 in type 1 HML-2, and the type 1 instead have been shown to express the alternatively spliced ~9 kDa protein Np9 [121, 124] As a reference, the putative splice sites and ORF for each type is labeled in Figures 3.3 and 1.14. The majority of type 1 proviruses encode a premature stop codon just upstream of the Δ292 start, and thus express just the first ~15 amino acid residues of the Env polyprotein [122].

We analyzed the frequencies of type 1 and 2 proviruses within our dataset to generate a more accurate representation of each type within the HML-2 group. Using the full-length alignment generated from the total number of identified HML-2, each provirus was inspected for the presence or absence of the 292 bp *pol-env* deletion diagnostic of type 1 structure (also refer to Figure 3.9). Elements that were observed to share either the upstream or downstream boundaries of this deletion were assigned a type, although for the majority the nucleotide sequence was intact across the site. For example, the 10q24.2 provirus has suffered a deletion that includes the 5' edge of the 292bp deletion, however the internal sequence of the provirus extends over the 3' portion of the type 1 site, and so was interpreted as a type 2 element (Figure 3.3 B). From analysis of the structure across this site, we were able to assign 73 elements to either subtype; the remaining 16 proviruses contained deletions that spanned the type 1 site and were excluded here. The results are summed in Table 3.2.

Of the 73 proviruses differentiated by structure, a majority of 53 were type 2 (~73%), while the remaining 20 were type 1 (~27%) (respectively, solid gray or black triangles in Figure 3.6 A). Prior estimation of the type 1 frequency within HML-2 was from the analysis of a subset of 35 HML-2 elements in 2003, again in a report with most

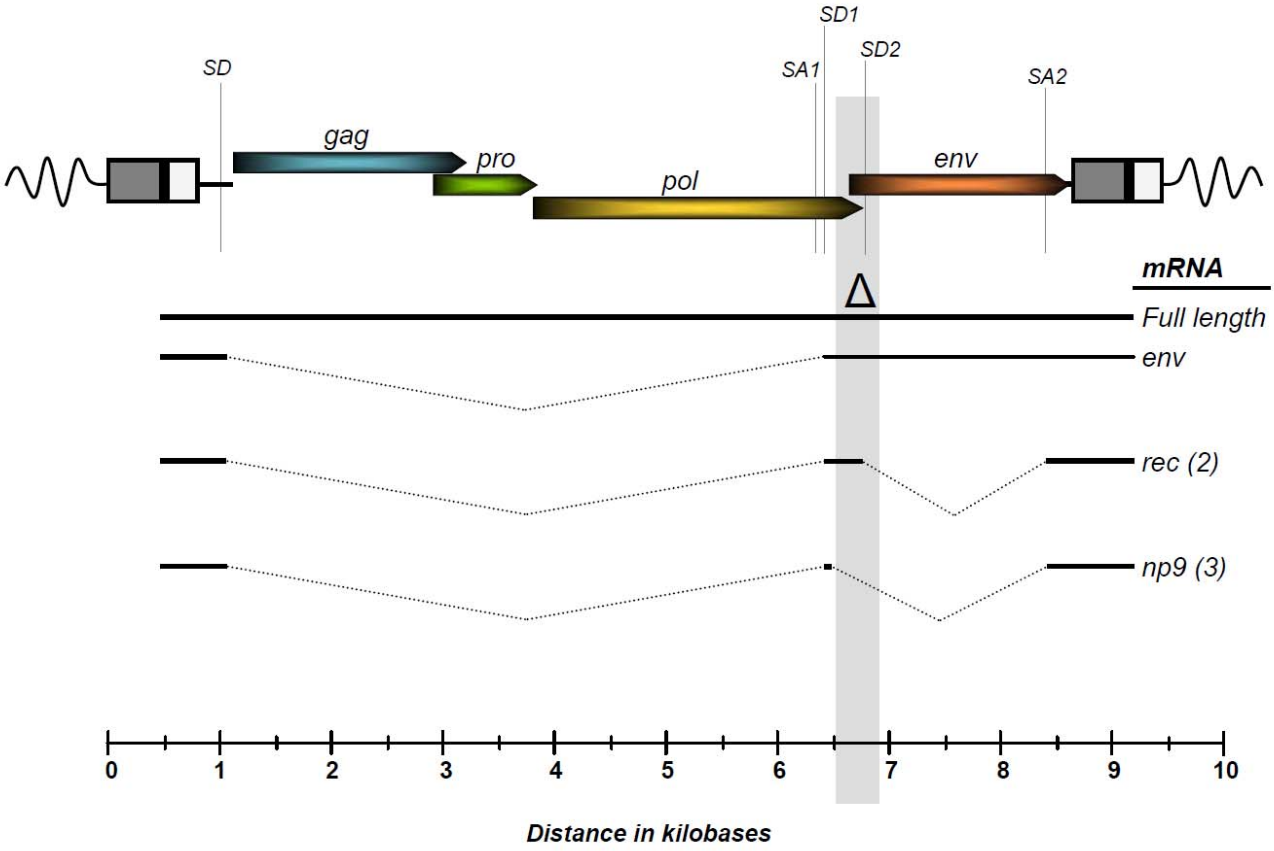


Figure 3.9. Scaled representation of HML-2 subtypes.
Upper: The structure of an integrated HML-2 provirus is shown with the essential genes labeled. Colors used for each reading frame and for the LTRs is as in previous figures. The positions of splice donor (SD) and acceptor (SA) sites are indicated with a line and labeled as such. The 292 bp deletion diagnostic of type 1 HML-2 is inferred by grey shading and the Δ symbol. Wavy lines represent the flanking regions of the host genome. Lower: The full-length and spliced mRNAs are shown for proviruses of either HML-2 subtype. Env is spliced from SD to SA1 in both subtypes. Rec (type 2) is spliced from the SA1 to the SD2 sites; Np9 (type 1) is spliced from the SA1 to SD1 sites. All transcripts are spliced to SA2. The reading frames for *rec* and *np9* transcripts are indicated in parentheses.

attention placed to the human-specific and most recently formed HML-2 integrations [94]. From the referenced collection of HML-2 proviruses at the time, about 44% of those able to be conclusively classified were identified as type 1. Thus, it appears that the type 1 allele is present at a frequency at least 10-13% lower among the whole of HML-2 than previously thought.

As >50% proviruses belonged to the LTR5-Hs, we asked whether the distribution of either subtype were also skewed within a particular LTR subgroup. For these purposes, the subtype classification for each provirus was superimposed onto the representative phylogenies of each HML-2 subgroup (Figure 3.6 A). Indeed, this representation revealed a distribution in which all type 1 elements were represented within the LTR5-Hs subgroup. The node common to the type 1 proviruses implies that the original Δ292bp allele emerged within the LTR5-Hs following divergence from the B subgroup, but prior to the formation of human-specific elements (arrowed in Figures 3.5 and 3.6 A). The relative times of formation for most of the proviruses near this node have been characterized by PCR of orthologous sites within multiple primates [64, 65, 94], and through *in silico* comparison of genome sequence from the available species [71]. With specific reference to the species distributions provided in these reports, one can infer that the type 1 allele originated prior to the common ancestor of humans and great apes.

The type 1 elements had a higher frequency than type 2 within the LTR5-Hs subgroup, with 20 (~54%) type 1 elements, and 17 (~46%) type 2, excluding the 8 proviruses with large deletions spanning the Δ292 site. Further inspection of the type 1 distribution within the LTR5-Hs tree revealed that, of the 19 analyzable human-specific proviruses, the subtypes 1 were present at roughly equal frequencies, with 9 (~47%) type

1 and 10 (~53%) type 2. Interestingly, the type 1 elements were less frequent among the polymorphic LTR5-Hs, with 4 (~36%) being type 1 and 7 (~64%) type 2.

Table 3.2. Summary of type 1 and type 2 prevalence among the HML-2 LTR subgroups.

		HML-2 subtype [a]		Not analyzed for subtype [b]
	∑	Type 1	Type 2	
HML-2	89	20	73	16
LTR5B	23	0	17	6
LTR5A	21	0	19	2
LTR5-Hs	45	20	17	8
H.s.-specific [c]	22	9	10	3
Polymorphic [d]	11	4	7	0

[a] Each element sorted by structure of the presence/absence of the Δ292bp of type 1.

[b] HML-2 for which the site containing the Δ292bp was no longer present by deletion.

[c] HML-2 specific to humans; *H.s.* refers to *Homo sapiens*.

[d] Human-specific sites detected with multiple alleles in humans.

A relatively small number of proviruses were available for such analysis, however a few trends do emerge. In general, type 2 elements are more commonly observed among the most recent integrations, however represent less than half of the typeable LTR5-Hs subgroup. Following its appearance within the LTR5-Hs elements, type 1 is the only observed subtype until the formation of human-specific proviruses, but is not monophyletic within the clade [94, 115]. However, as at least 8 proviruses were not able to be analyzed by subtype, it is quite possible the deletion arose prior to its apparent clade, and so this is considered a tentative time of the type 1 emergence. The patterns and

frequencies of the type 1 elements within the tree, and their lower presence within the newest HML-2 germline insertions may reflect a tendency within this group for subtype conversion, and indicate a relatively high level of exchange of the type 1 deletion multiple times in the germline. This explanation, and alternative hypotheses, are given further attention in the Discussion section of the chapter.

Investigation of HML-2 proviruses with evidence of duplication

As described briefly, phylogenetic analysis of the LTRs belonging to individual proviruses may be used to infer genome-level recombination and gene conversion events. In such a tree, aberrant groupings of cognate 3' and 5' LTRs may also signify proviruses that have been subject to other rearrangements. For instance, stably integrated proviruses involved in duplication(s) will also exhibit unexpected clustering of LTRs. Because they are identical upon integration, the 3' and 5' LTRs of that provirus are more similar to one another than to any other LTRs [63-65, 71]. In the event a provirus is duplicated in the context of the surrounding region, the LTRs will share most identity by virtue of the copied element, such that the paralogous 5' LTRs and 3' LTRs will group together (Figure 3.10 A and B). Another prediction of such an event is that the target site duplications (TSDs) of each copy remain identical (Figure 3.10 C). Evidence of both attributes may infer proviruses that benefited from an increase in copy number in the absence of exogenous activity.

Examination of the LTR phylogenies revealed multiple clades with deviated groupings of the 5' and 3' LTRs in the predicted topology for elements having undergone duplication within the genome (indicated in Figure 3.6 B and C). This specific clustering pattern was observed within three well-supported clades over the LTR5A and B subgroup trees, but was absent from the representative LTR5-Hs tree. Also, the branching patterns for each cluster were in complete agreement between the LTR-based and gene-based trees, giving further support to their shared ancestry (data not shown).

The contribution of host duplication events to copy number in the HML-2 (in addition to other HERVs) has been mentioned in passing, with specific reference to just

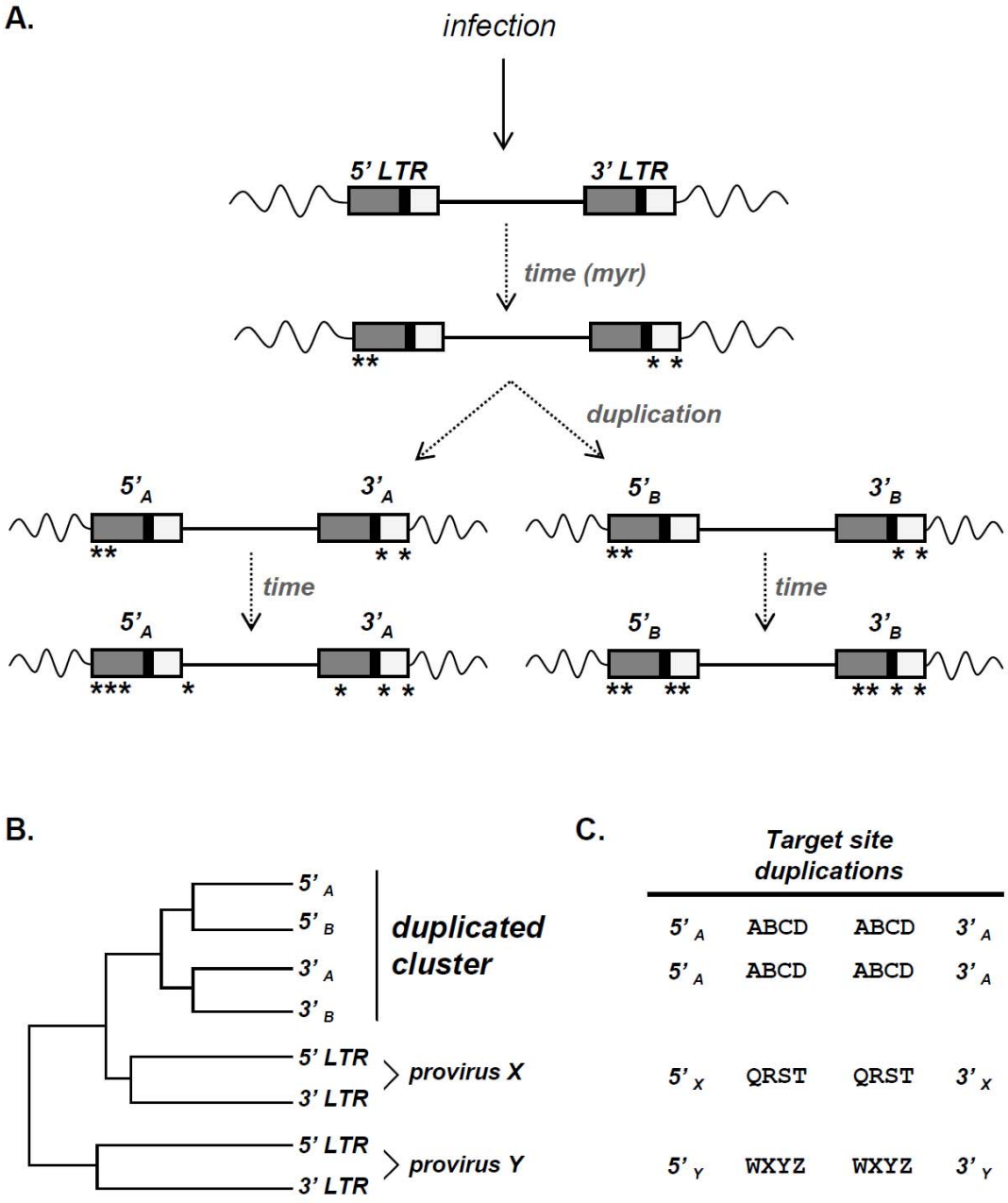


Figure 3.10. Representation of LTR evolution and the detection of duplication.
A. An endogenous provirus is shown following infection of the germline. Mutations are indicated by asterisks below wither LTR over time (in millions of years). A duplication event is indicated by split arrow. Following duplication, mutations are shown to accumulate independently over time.
B. LTR trees indicate pairing of the 5' and 3' LTRs of four different elements over time. The unexpected clustering indicative of duplication is labeled.
C. The representative target site duplications are indicated at right, shown to be identical for the duplicated provirus.

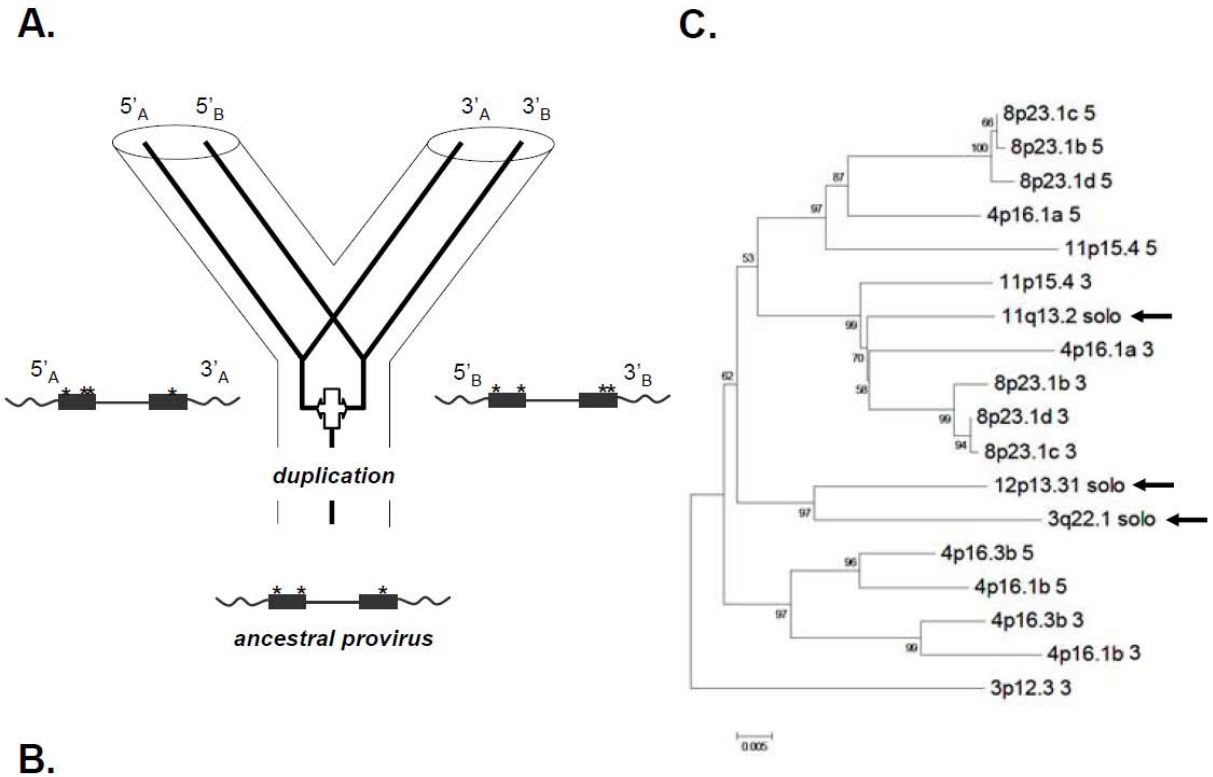
one such cluster of elements (within the LTR5A and described in further detail below). In only two or three reports have these elements been noted, in passing and for exclusionary support of other data [93, 120], and it has been suggested that copy number affects from host duplication have probably been limited [38]. The analyses below provide evidence of multiple unique clusters within the LTR5A and LTR5B subgroups that arose during multiple large-scale duplications within the genome, and suggest such events have led to about a third of the representative LTR5A, and half of the LTR5B.

Support for segmental duplication within the LTR5A subgroup

The largest clade within HML-2with evidence of duplication was in the LTR5A with two lineages represented by the 11p15.4, 4p16.1a, and 8p23.1b, c, and d proviruses, and the 4p16.1b and 4p16.3b proviruses (upper and lower in Figure 3.6 B). Collectively, the proviruses making up both groups had several features in common, such as short insertions, deletions and single base changes relative to K113 (also refer to Figure 3.3 B). Within the larger group, two lineages with deviated clustering were observed from a single node, one giving rise to the 5' LTRs, and the other giving rise to the 3'LTRs for each of the five proviruses. The proviruses making up the group shared a high level of nucleotide identity (96%), particularly among the elements at 8p23.1 (>99% identity). Similar to the LTR clustering for these proviruses was the pattern observed for the smaller group containing the 4p16.1b and 4p16.3b proviruses. Likewise, the 4p16.1b and 4p16.3b proviruses were highly similar, with close to 98% identity. Branching most closely was the 3p12.3 provirus, which shared ~94% identity to the two groups.

The TSDs were examined for each provirus, including the single LTR from the outgroup provirus to the cluster, 3p12.3, to investigate the putative events leading to the observed phylogeny (bolded in Figure 3.11 B). For these purposes, 30 bp of cellular sequence flanking the 5' and 3' LTRs for each provirus were aligned with reference to the LTR edges. Comparison of the TSDs from each provirus confirmed each pair was intact, and revealed the flanking genome sequence was identical between proviruses, both indicators of shared ancestry. In contrast was the single 3p12.3 3'LTR, which shared neither TSD nor flanking sequence, suggesting its unique insertion; however given the lack of a combined analysis with the 5'LTR, the possibility of gene conversion or recombination involving this provirus is difficult to exclude.

As a primary characterization of the host sequence flanking each provirus, we extracted and BLAT-searched the Hg19 build with ~10kb of genome sequence in either direction of each provirus. The search was later performed using a window of up to 200kb, for reasons discussed in the analyses below. Best scores for each search were from hits to the DNA regions flanking the other seven proviruses (from ~95% to >99%). Also identified were a few DNA segments of similar or smaller size but without associated proviruses, likely from older derived copies given their lower sequence identity (<94% identity to the query). Interestingly, a chromosomal region within 100kb from the 3p12.3 provirus was positively hit in each search, though not the 3p12.3 provirus itself. This observation, and the additional hits described below, prompted a more detailed examination of the genome regions surrounding these elements.



B.

```
                                    < 5' LTR                                  3' LTR >
8p23.1c          TGGGCTCTTTATTTG TGTAGGGAAAAGAA...AGGCCACCCCTTCA ATTTGGTGACACATA
8p23.1d          TGGGCTCTTTATTTG TGTAGGGAAAAGAA...AGGCCACCCCTTCA ATTTGGTGACACATG
8p23.1b          TGGGCTCTTTATTTG TGTAGGGAAAAGAA...AGGCCACCCCTTCA ATTTGGTGACACATG
11p15.4          TGGGCTCTTTATTTG TGTAGGGAAAAGAA...AGGCCACCCCTTCA ATTTTGTGACACATA
4p16.1a          TGGGCTCTTTCATTTG TGTAGGGAAAAGAA...AGGCCACCCCTTCA ATTTGGTGACACATA
4p16.1b          TGGGTTCTTTTATTTG TGTAGGGAAAAGAG...AGGCCACCCCTTCA ATTTGGTGACACATA
4p16.3b          TGGGTTCTTTTATTTG TGTAGGGAAAAGAG...AGGCCACCCCTTCA ATTTGGTGACACGTA   provirus

12p13.31 solo    TGGGCTCTTTATTTG TGTAGGGAAAAGAA...AGGCCACTCCTTCG ATTTGGTGACACATA
3q22.1 solo      TGAGCTCTTTATTTG TGTAGGGAAAAGAG...AGGCCAACCCTTCA ATTTGGTGACACATA
11q13.2 solo     TGGGCTCTTTATTTG TGTAGGGAAAAGAA...AGGCCACCCCTTCA ATTTGGTGACAAATA   solo
3p12.3                                           AGGCCACCCCTTCA TATGAGGAATTCAGA

3q21.2           GCAAGATCTGGGCCC TGTGGGGAACAGCA...AACCCACCCCTACA GGCCCAGAGCACAAT   K(I)

5q33.2           TATAAAATCATTACT TTTAGGGAAAAGAA...AGGCCACCCCTTCA ATTACTGAGCTATTT
6p11.2           ATAGTTACTTCTTGT TGTAGGGAAAAGAA...AGGCCACCCCTTCA CTTGTCTTCTGCTAG
9q34.11          ATCCTCCTTCTTCAG TGTAGGGAAAAGAA...AGGCCACCGCTTCA CTCAGCTTCCCAAAG   other
```

Figure 3.11. Germline duplications of proviruses within the LTR5A subgroup.

A. Shown is a diagram of the predicted phylogeny for a duplicated provirus. At bottom, an integrated provirus is shown, which is subsequently duplicated. The single branch leading to the duplication event represents the LTR divergence after insertion, but prior to duplication. Accumulated changes independent to either LTR are asterisked. After duplication, the 5' and 3' LTRs of each proviral copy will be more similar to one another, than the 5' and 3' LTRs of a single copy. Barring recombination or gene conversion, the duplicate 5' and 3' LTRs will cluster in a tree.

B. A nucleotide sequence alignment of the junction sites for elements associated with the LTRA duplicates, indicated by the LTR5A duplicated proviruses (pro), solo LTRs (solo), the K(II) element from LTR5-Hs, and representative LT5A without evidence of duplication (other).

C. A neighbor-joining tree constructed from the 5' and 3' LTRs from 10 LTR5A elements with evidence of duplication. Arrows indicate the position of 3 solo LTRs identified by BLAT search of the genome sequence flanking each duplicated provirus. Tree construction was as described for the LTR5A subgroup phylogeny in Figure 3.6 B, and rooted with the 3p12.3 single 3' LTR.

Three separate sites consistently hit were colinear to the queried sequence, but found to contain solo LTRs belonging to the LTR5A subgroup, specifically at 3q22.1, 11q13.2, and 12p13.31. The TSDs flanking each solo LTR were found to be intact and identical to those of the LTR5A cluster (also aligned in Figure 3.11 B), suggesting their ancestry within the same lineage as the LTR5A duplicates. To examine this possibility, we analyzed the relationship of each solo LTR to the proviral LTRs of the subgroup A duplicates in a neighbor-joining tree rooted with 3p12.3 (Figure 3.11 C). All three solo LTRs clustered within the LTR5A tree. The 11q13.2 solo LTR was observed basal to the 3'LTRs within the larger proviral lineage, whereas the 12p13.31 and 3q22.1 solo LTRs formed sister taxa on a branch intermediate to either proviral lineage (each is arrowed in Figure 3.11 C). Given this arrangement, the 11q13.2 allele may have arisen via independent solo LTR formation, while the latter elements were likely copied from a single ancestor.

An explanation for the observed pattern at each site is segmental duplication. To determine the extent of segmented genome surrounding the LTR5A duplicates, we expanded the search to a ~200kb window, and performed a detailed comparison between the DNA flanking each site. Indeed, concentrated stretches of similar sequence were observed from about 50kb to 100kb from the boundaries of each provirus, especially in the context of nearby REs. Within the 200kb window, the locations of two proviruses were resolved, at 8p23.1c and d. Using these sites as anchors, we mapped the remaining LTR5A duplicates. The representative map is shown in Figure 3.12 A, in which the shared position of each element has been superimposed onto a portion of the 8p23.1 chromosomal locus that corresponds to its native segment (windows of individual

chromosomal regions from the UCSC Genome Browser are detailed in Appendix A). Many of the segments exceeded 100 kb, and a few of the largest had lengthy regions of unaligned gaps in the genome (for example indicated in Figure 3.12 A). At least within the 8p23.1 chromosomal region, the gaps have been analyzed in fosmid-based end-sequencing studies, and shown to have highly variable structures in both length and RE content [192]. Therefore, we could only speculate on the true span of these segments. Expanding the search window over the 3p12 region of the same duplicated segment (described above) included the 3p12.3 insertion site (arrowed at right in Figure 3.12 A), of interest as the provirus serves as the outgroup to the LTR5A duplicates, but appears to be a unique site by virtue of location and intact TSDs.

Notably, two more sites were identified within the same segmented regions that had apparent unique HML-2 integrations at locations nearby the LTR5A copy site. The first was returned in the original BLAT searches, a provirus at 3q21.2, again colinear with >80kb of the same flanking segment (also arrowed in Figure 3.12). This site was in exception to the pattern described above, as a unique insertion of an LTR5-Hs provirus was observed. The TSDs were intact, suggesting its independent origin. In fact, the chromosomal coordinates confirmed it to be the HERV-K(I) provirus, a relatively well-studied HML-2 first described in 2001 [188]. In order to map the LTR5-Hs and LTR5A elements in the context of the surrounding cellular sequence, we compared the relative base positions of the 3q21.2 and LTR5A proviruses. Quite surprisingly, the 3q21.2 provirus was located within 5kb directly upstream from that shared among the LTR5A duplicates (asterisk in Figure 3.12).

A similar situation was encountered for the third identified HML-2, which we identified within a region of tandem segment adjacent to the duplicon containing the 8p23.1b provirus. This particular site has been labeled within the window as the 8p23.1b LTR5A element for reference in Appendix A. We confirmed that the provirus was in fact the 8p23.1 LTR5-Hs element, or more specifically, HERV-K115. The distance between the 8p23.1b element and the K115 provirus was around 700kb, flanking an intermediate gapped region of unaligned chromosome. K115 is situated near the middle of, but does not border, a large inversion of two duplication segments in the opposite orientation. Due to the ambiguity of the unaligned regions, we were unable to map its distance to the nearest set of segmental core REs. The association of the additional 3q21.2 and K115 elements to the same types of segmental duplicons as the LTR5A is speculated further in the discussion section of this chapter.

Taken together, these findings suggest the stable formation of an endogenous HML-2 provirus in a genome region that was subsequently amplified in the germline, with the effective increase in copy number and genome expansion evolution. The duplicated elements collectively account for 7 of 21 proviruses belonging to the LTR5A subgroup, suggesting that at least a third of these elements were formed without exogenous activity. Given the large segments of shared sequence, homogenization by gene conversion is difficult to conclusively dismiss, and could have affected or contributed to the LTR5A copies analyzed here. However support against gene conversion at these sites is offered in the phylogeny of these elements, whose LTRs cluster in the topology predicted for duplication. Regardless, these results indicate the

relatively rare germline activity of the LTR5A subgroup, particularly in comparison to the LTR5-Hs elements.

Inferred reconstruction of LTR5A genome expansion via a specific group of recent segmental duplications

Given the extensive genome region shared between each of the LTR5A duplicates, their detectable copy number from the published database, and phylogenetic support in clustering of LTRs, we asked whether the chromosomal regions involved had evidence of a propensity for inter- or intrachromosomal rearrangements. Comprehensive literature searches presented the recent reports of a group of 'composite' segmental duplications, whose chromosomal distribution was highly similar to that observed for the LTR5A duplicates. Referred to as 'DA' (after the Chinese character for 'large'), the segment is present at about 20 copies, specifically on chromosomes 3, 4, 7, 8, 11, and 12, within or near regions associated with the LTR5A cluster (also see Table 3.3) [193].

By comparing the genome regions in the context of each LTR5A duplicate (including proviral and solo LTR elements), we identified a region shared by all, or nearly all duplicated loci, that exhibited essentially the same 'core' REs and structural ordering, shown schematically in Figure 3.12 B and C. Specific denoting features included, in the downstream direction of the LTR5A copy site, the insertions of a HERV-E, HERV-S71, and LTR5B (solo) interspersed with *Alu* and L1 elements. The HERV-E provirus contained a secondary HERV-H integration in some, but not all, cases. At the extreme flanks of the largest duplications (presumably the most intact) were repeated regions of additional HERV-E elements with satellite DNA. Comparison with the DA

Table 3.3. Characteristics of the LT5A duplicates and their associated DA segments.

LTR5A locus	Start (bp)[a]	DA locus (chr-Mb)[c, d]	DA type[d]	Oldest sp. (PCR) [d, e]	notes
3p12.3	75600465	3p-75	I	Gb	Unique LTR5A integration
3q22.1	129842183	3q-127	I	O	Solo at LTR5 site
4p16.1a	9123515	4p-9	II/III	G	Next most recent to 8p23.1 loci
4p16.1b	9659580	4p-9	II	G	HERV-E only
4p16.3b	3979051	4p-4	II	G	HERV-E only
8p23.1b	8054700	8p-7-8	II/III	G	Adjacent DA segment is with 8p23.1a (K115)
8p23.1c	12073970	8p-11-12	II/III	H, C?	Flanks 8p23.1d DA
8p23.1d	12316492	8p-11-12	III	H, C?	Most recent DA, with specific 8p insertion
11p15.4	3468656	na	na	na	No mention in DA reports
11q13.2	3577183	11q-6-7	II/III	H, C?	Solo at LTR5 site
12p13.31	8417281	12p-8	III	G	Solo at LTR5 site

[a] Chromosomal position is with reference to the Hg19 genome build (February, 2009).

[b] Estimated from 5'/3' LTR comparisons and normalized to a substitution frequency of 2x0.13% per mya, from the method described by Lebedev, *et al*., 2000 [189]. Same as in Table 3.1.

[c] Chromosomal position is with reference to the Hg18 build (March, 2006).

[d] From Ji, X., *et al*. 2008. DA and Xiao – two giant and composite LTR-transposon-like elements identified in the human genome. *Genomics* 91:248.; Li, X., *et al*. 2009. Stepwise evolution of two giant composite LTR-transposon-like elements DA and Xiao. *BMC Evolutionary Biology* 9:128. Included with permissions.

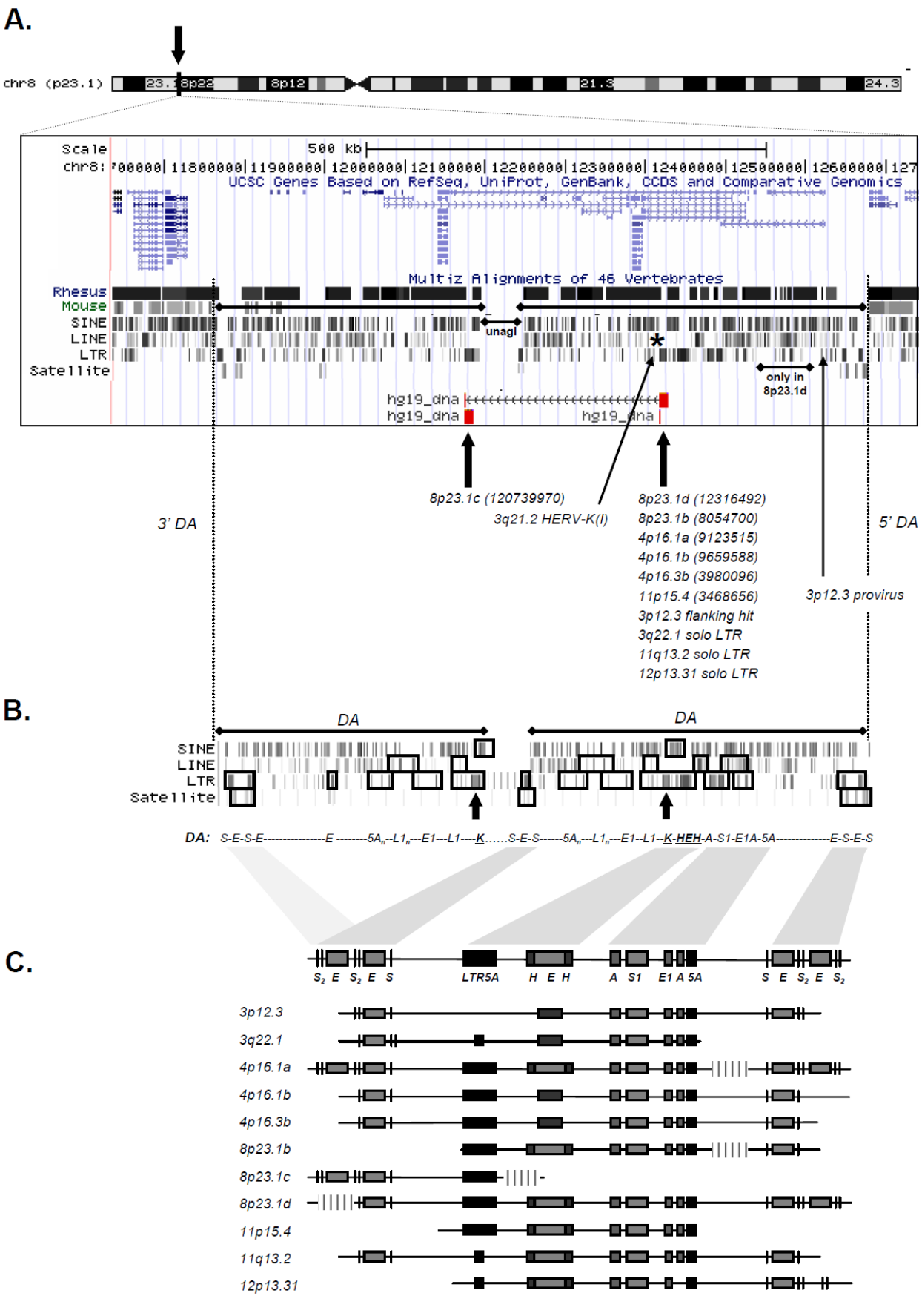
[e] Abbreviations as follows: H, human; C, chimpanzee; G, gorilla; Gb, gibbon; O, orangutan.

B. A mapped derivation of a DA 'core' repetitive element (RE) sequence with reference to the 8p23.1 locus. Arrows show positions of 8p23.1c and d, as in (A). RE classes are indicated at left; those shared in most HML-2-containing DA are outlined within the map. The derived core is below: *S*, satellite; *E*, HERV-E; *5A*, LTR5A solo LTR; *L1*, LINE1; *E1*, ERV1; *K*, duplicate LTR5A position; *S1*, HERVS71. Shaded boxes are used to denote identified 'core' RE used for comparative analysis. Light grey boxes were used to

indicate the edge of the minor DA (containing the ‘c’ LTR5A copy) located in tandem with the major DA (containing the ‘d’ copy) at the 8p23.1 locus.

C. Top: The core RE derived from HML-2-containing DA segments. Bottom: A schematic representation of each of the 10 LTR5A duplicated elements in the context of the surrounding DA core RE.

D. A proposed model for the genome expansion of a subset of LTR5A as mediated by DA-specific segmental duplication events. Dotted lines represent duplications leading to the 8p23-containing LTR5A cluster; the 4p16.1b and 3b cluster, and the 3q22.1 and 12p13.3 solo LTRs. Putative events in solo LTR formation within the LTR5A duplicates, and events leading to increased complexity in DA segments are labeled. Asterisks indicate events that either involve solo LTR formation at the LTR5A site, or are not resolved between the data. The light arrow from 11q13.2 to 4p16.1a indicates possible copying prior to solo LTR formation.


A.
chr8 (p23.1)
23.1
8p22
8p12
21.3
24.3
Scale
500 kb
chr8:
11800000
11900000
12000000
12100000
12200000
12300000
12400000
12500000
12600000
127
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
SINE
LINE
LTR
Satellite
unagl
only in
8p23.1d
hg19_dna
hg19_dna
hg19_dna
8p23.1c (120739970)
3q21.2 HERV-K(I)
8p23.1d (12316492)
8p23.1b (8054700)
4p16.1a (9123515)
4p16.1b (9659588)
4p16.3b (3980096)
11p15.4 (3468656)
3p12.3 flanking hit
3q22.1 solo LTR
11q13.2 solo LTR
12p13.31 solo LTR
3p12.3 provirus
3' DA
5' DA
B.
DA
DA
SINE
LINE
LTR
Satellite
DA: S-E-S-E----------E --------5A$_n$--L1$_n$---E1---L1----K.....S-E-S------5A$_n$---L1$_n$---E1--L1--K-HEH-A-S1-E1A-5A----------E-S-E-S
C.
S$_2$ E S$_2$ E S
LTR5A
H E H
A S1
E1 A 5A
S E S$_2$ E S$_2$
3p12.3
3q22.1
4p16.1a
4p16.1b
4p16.3b
8p23.1b
8p23.1c
8p23.1d
11p15.4
11q13.2
12p13.31

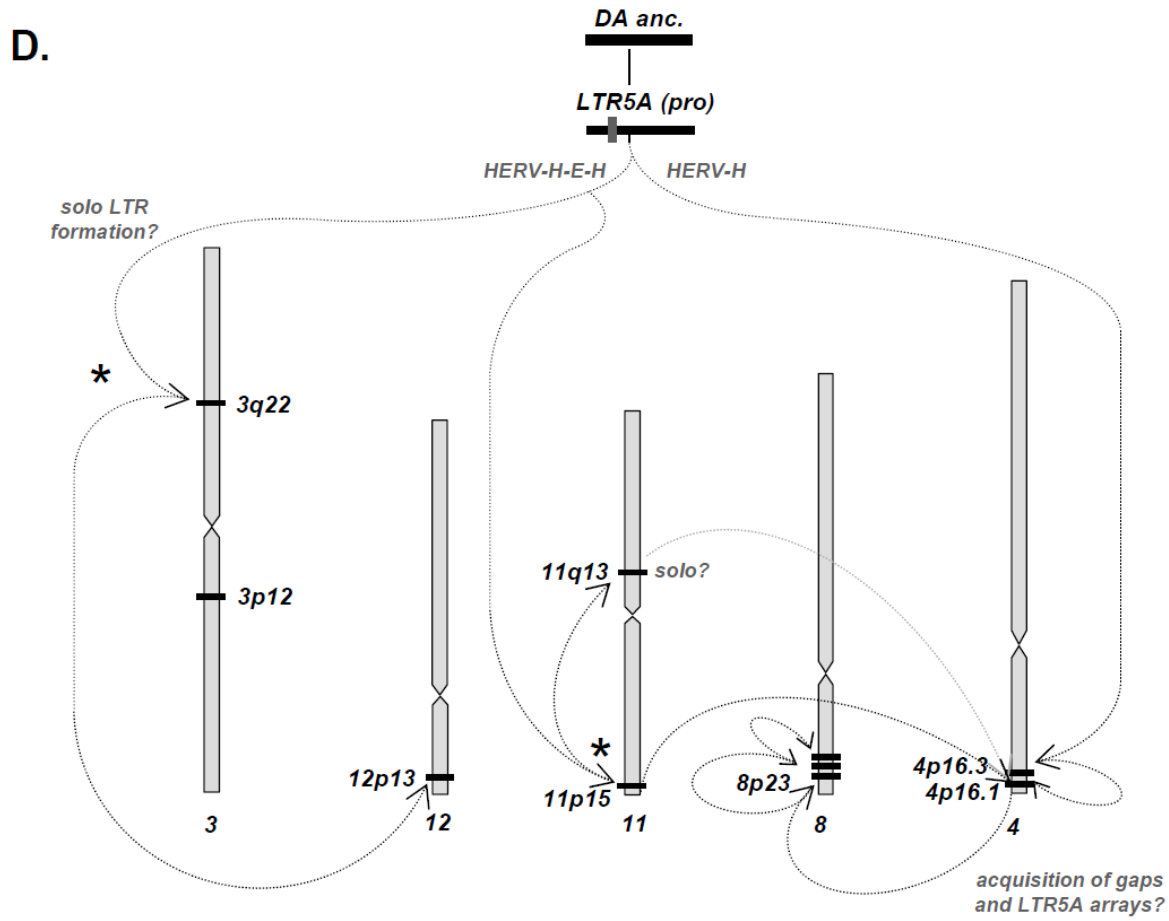


Figure 3.12. LTR5A genome expansion was mediated by segmental duplication within the large 'DA' composite-like genome regions.

A. Shown is a snapshot of chromosomal position 8p23.1 from the UCSC Genome Browser displaying a 'DA' segmental duplication. Above is relative location on chromosome 8. Within the window, two duplicated segments are marked with a knobbed line; the edges of the duplication are delineated with dotted lines. The central region contains an unaligned gap, also marked with knobbed line. The window includes the 8p23.1c and d elements, positions indicated with block arrows. At the 8p23.d locus, the remaining LTR5A duplicates have been indicated by chromosomal locus and reference start position in parenthesis. Also shown are the relative sites of the 3p12.3 (arrow) and 3q21.2 proviruses (arrow, asterisk).

'core' revealed the same, or nearly the same pattern of RE motifs and ordered structure [193], with the noted exception of the presence of the LTR5A site, which was not mentioned in previous reports. Mapping the core REs in the context of each LTR5A copy revealed that most LTR5A were contained within apparent full-length DA segments (6 of 10) (Figure 3.12 C). Few LTR5A-containing duplicons shared all core REs.

DA segments have previously been divided into three classes by virtue of RE content, the most basic feature being the HERV-H insertion downstream of the LTR5A copy site [193, 194]. Most basically, the type II elements carry the HERV-H insertion, and type III carry the added integration of the HERV-E [193, 194]. Characterization of the species distribution [194] has shown that, while type I DA are detected in orangutan, the types II and III are restricted to the humans, chimpanzees, and gorillas (also refer to Table 3.3). Type III are the most recently formed DA, and possess an additional internal duplication derived from the 8p23 region, reminiscent of the 8p23.1d LTR5A-containing duplicon (indicated by knobbed line in Figure 3.12 A). Using this information, in combination with our phylogenetic analysis (Figure 3.11 C) and the identification and mapping of core REs to each LTR5A duplicate, we attempted to reconstruct the duplication events leading to the presently observed LTR5A distribution in the genome.

A putative model is provided in Figure 3.12 D for the events contributing to the LTR5A copies. Following the primary LTR5A insertion, a duplication generated the segments leading to the 4p16.1b and 3b, and 8p23.1 clusters. This specific split is supported from the phylogeny of these LTR5A, in addition to their RE core content. Prior to the amplification of the 5 segments within the 8p23 LTR5A cluster, we can speculate another duplication leading to the formation of the 3q22.1 and 12p13.31 solo LTRs. We

can predict the HERV-E integration into the DA core occurred prior to this event, as it is shared between these solo LTRs and the elements within the 8p23.1-containing cluster, and was likely after the gorilla-orangutan split, as suggested by PCR distribution by others (also see Table 3.3) [194]. The remaining modeled events are mostly inferred from phylogeny, with the final generation of the segments at 8p23.1.

Because these elements have been restricted to the germline, the branching patterns and lengths are reflective of the evolution between the LTRs from each copy while stably integrated. Any change between LTRs of individual copies prior to duplication will be preserved in a subsequent copy. Thus, differences in the 5' and 3' LTRs of a single copy should be more reflective of the ancestral element, a prediction from which is an overestimation of the ages of recently formed copies. For example, the branching observed for the 8p23.1 elements suggested little change between each duplication, and indeed the copies are >99% identical, however each is estimated to have formed ~20 mya (also refer to age vs. changes between the LTRs for these copies in Table 3.1). The high identity, and little change suggests some of these duplications may be specific to humans.

At least 19 of 20 DA were reported in the chimpanzee, based on PCR and *in silico* data [194], however an 'empty' site for the remaining DA was not confirmed. Interestingly, the corresponding region was within the 8p23.1 locus. We repeated the BLAT using the most recent chimp genome build (Oct. 2010, 2.1/panTro3), however hits were obtained for each site, one of which included an unaligned contig to the same chromosome (data not shown). With a lack of conclusive data from at least the human and chimpanzee, it remains unclear whether any DA containing the LTR5A were formed

after speciation. Regardless, the collective analyses of this duplicated cluster within the LTR5A subgroup confirms that at least a third of the LTR5A proviruses are accounted for within the duplication. Further, half of the DA segments share the LTR5A insertion within their core, and appears to be present in only the most recently evolved type II and III DA copies, suggesting the LTR5A element is a part of the core RE motif belonging to these specific DA classes.

Analysis of duplications within the LTR5B subgroup

Within the LTR5B subgroup phylogeny, there were two clusters observed with branching reminiscent of origins via duplication, respectively located at chromosomal locations 1p36.21 and Xq28 (upper and lower bars in Figure 3.13 C). An additional two sites, containing the 15q25.1 and Yq11.23 LTR5B (refer to Figure 3.3 C for inferred structure), also showed evidence of duplication, supported in both *pol* and *env* trees at bootstrap values at or near 100 (data not shown), but were not investigated further here as each was no longer associated with its LTRs. Of note, two proviruses at 1p36,21, and the full-length Xq28b have been described, however their corresponding duplicated copies have not been previously reported, nor have the internal duplications on chromosomes 15 and Y mentioned above. Thus, this is the first reported evidence of segmental duplications of HML-2 within the LTR5B.

The duplicates from the 1p36.21 and Xq28 clusters were analyzed in a similar fashion as described above for the LTR5A duplications. Briefly, the TSDs for each provirus were examined, and similarly, each was found to be intact (Figure 3.13 A). Proviruses within each cluster were also observed to share large segments of flanking

cellular sequence, as indicated from BLAT analysis. Consecutive BLAT searches were performed as described above, initially with ~10kb and ~25kb flanking sequence, and the search window later expanded to a ~200kb span. The 1p36.21 LTR5B proviruses shared at least ~200kb of surrounding genome sequence, and the Xq28 copies share at least ~50kb of genome region distinct from 1p36.21 (Figure 3.13 B and C). There was a striking level of sequence identity to the other(s) within its group: both Xq28 copies, and

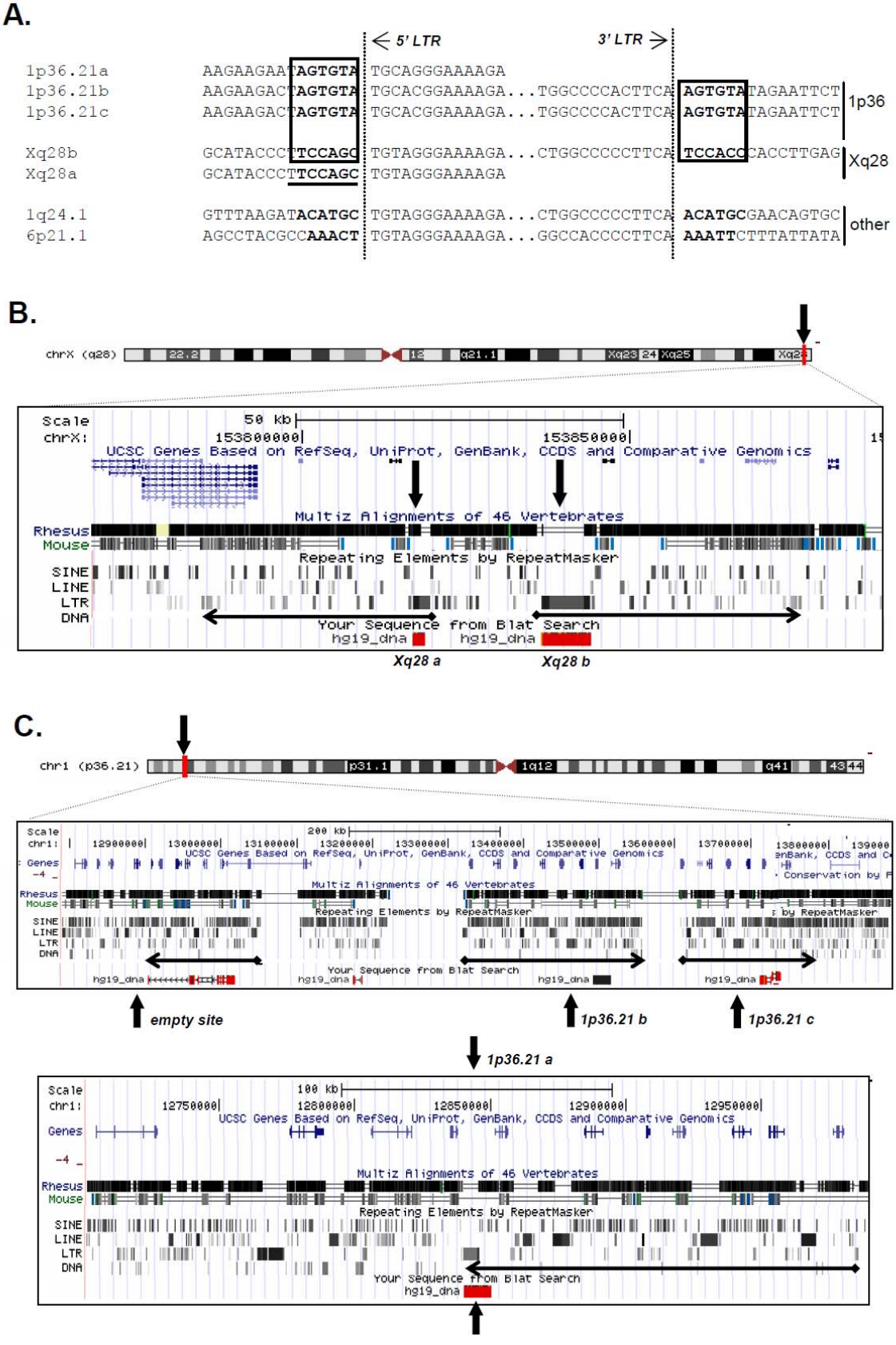

A.
5′ LTR
3′ LTR
1p36.21a AAGAAGAATAGTGTA TGCAGGGAAAAGA
1p36.21b AAGAAGACTAGTGTA TGCACGGAAAAGA...TGGCCCCACTTCA AGTGTA TAGAATTCT
1p36.21c AAGAAGACTAGTGTA TGCACGGAAAAGA...TGGCCCCACTTCA AGTGTA TAGAATTCT
1p36
Xq28b GCATACCCTTCCAGC TGTAGGGAAAAGA...CTGGCCCCCTTCA TCCACC CACCTTGAG
Xq28a GCATACCCTTCCAGC TGTAGGGAAAAGA
Xq28
1q24.1 GTTTAAGATACATGC TGTAGGGAAAAGA...CTGGCCCCCTTCA ACATGC GAACAGTGC
6p21.1 AGCCTACGCCAAACT TGTAGGGAAAAGA...GGCCACCCCTTCA AAATT CTTTATTATA
other
B.
chrX (q28)
Scale
chrX:
50 kb
153800000
153850000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Your Sequence from Blat Search
hg19_dna
Xq28 a
Xq28 b
C.
chr1 (p36.21)
Scale
chr1:
200 kb
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Multiz Alignments of 46 Vertebrates
Repeating Elements by RepeatMasker
Your Sequence from Blat Search
empty site
1p36.21 b
1p36.21 c
1p36.21 a
100 kb
12750000
12800000
12850000
12900000
12950000

Figure 3.13. Evidence for two unique duplicated clusters within the LTR5B subgroup of HML-2 proviruses.

A. Nucleotide sequence alignment of the junction sites for LTRB duplicates, indicated by the two proviral clusters (labeled 1p36 and Xq28) and representative LT5B without evidence of duplication (other).

B. UCSC Genome Browser snapshot of the Xq28 segmentally duplicated region containing the Xq28a and b elements. Position of each element is labeled and marked with an arrow. Duplicated segment is indicated by knobbed line.

C. Upper: UCSC Genome Browser snapshot of 1p36.21 segmentally duplicated region. Three duplicated segments are marked with separate knobbed lines. Arrowed ends where included indicate orientation of the segment. Positions of the 1p36.21b and c proviruses are indicated as in (B). Lower: USCS Genome Browser window of the 1p36.21a truncated duplication. Duplicated segments and proviruses are indicated as in (B) and (C).

1p36.21 b and c having respective identities that exceeded 99.9%; the 1p36.21a copy had 98% identity to its b and c counterparts.

The best hit scores for each BLAT search were the remaining proviruses in each cluster. No hits were observed between clusters in LTR5B or to any of the LTR5A (ie, a search with the Xq28 flanking segment never returned a sequence at or near 1p36.21), indicating each subtree arose from independent genome locations (data not shown). Also, whereas BLAT searches with the segmental region flanking each LTR5A led to the identification of additional site that contained solo LTRs, no such regions were observed for the 1p36.21 and Xq28 loci. In fact, although each set of duplications containing the LTR5B shared high sequence identity and structured order, the respective segment structures appeared to be unique.

The 1p36.21 proviruses were represented in three copies; two were full length ('b' and 'c' copies) and the third was completely deleted for the 3' end ('a' copy; also refer to Figure 3.13 C). The full-length 1p36.21b and c proviruses were situated within a direct duplicated region of ~200kb that shared a highly similar structural order (inferred by arrow and orientation in Figure 3.13 C). BLAT of the 1p36.21b an c flanking regions did verify shorter (roughly ~10-30kb) stretches of similar sequence that were dispersed over a range of nearly 1Mb (data not shown), and led to the identification of an upstream duplication of the same size, but which was not associated with a provirus at the cognate site ("*empty site*" in Figure 3.13 C). This particular segment was inverted with respect to the segments containing the 1p36.21b and c elements. More than 1Mb upstream, the partial 'a' copy shared nearly the length of the 1p36.21b and c segments, however the segment border could be mapped to the provirus, such that the 3' truncation marked its

edge. A similar organization was observed for the LT5B elements located at Xq28, in which the ‘b’ copy was full-length and the ‘a’ truncated, its 3’ edge also denoting the border of its duplicon.

Collectively, the results of this analysis suggest the LTR5B elements, in addition to those belonging to LTR5A, have had some expansion in the germline in the absence of infection or movement by retrotransposition. Although the LTR5B elements have had fewer than the LTR5A, the pattern of duplications observed from the both the 1p36.21 and Xq28 copies have the interesting trend that, in each case a copy was truncated, and the truncation appeared to mark the duplication segment. Such a pattern was not observed for the LTR5A duplicates, although those particular proviruses appear to be restricted to a specific duplicon, whereas the 1p36.21 and Xq28 seem far less distributed through the genome. The mode of duplication could vary between sites, or possibly have effected the evolutionary activity of the described segments. Nevertheless, these results indicate that, for the fewer duplicated LTR5B proviruses, each lineage has implication of direct involvement of the provirus in at least two segmental duplication events.

A PCR analysis of each site in the representative genomes of primates closely related to humans could provide further insight to the events leading to the observed state at each site, and could perhaps allow an estimation of the tempo for duplication events within these particular genome regions. As demonstrated in previous work in our lab, direct comparison between homologous integration sites can be used to infer times of integration, and further sequence analysis of individual proviral sites can provide an indication of the evolutionary history for each element. A likely challenge to such a PCR

study is the basis on identical sites of integration within duplicated sequences up at least 40-50kb, preventing the distinction of orthologous and paralogous copies at each site. Other confounding factors could be the existence of additional copies from species-specific duplications, or homogenization between sites by gene conversions. Thus, a comparative study of the available published genomes may provide further insight into the evolution of these particular cluster of elements.

Discussion

Here we have presented the cumulative results of a genome-wide analysis of the currently described proviruses belonging to the HERV-K (HML-2) group. As a basis for the analysis, we generated a catalog of the HML-2 proviruses present in the human genome by mining the available databases and with reference to added literature. In doing so, we have identified 89 proviruses, adding nearly 30 previously undescribed elements closely related to this group of HERVs.

Because our search strategies included individual reading frames representative of the most conserved HML-2 (in all cases, K113 was queried, with supporting searches from the engineered K_{CON} and PHOENIX), elements with relatively high percent identity, but which have suffered large deletions were identified, but also identified were full-length or near full-length elements with lower percent identity to the complete K113 sequence that were indeed shown to belong to the HML-2. The final elements returned from each search had the lowest nucleotide identity to the queried provirus and the HML-2 group, and formed a clade distinct from the HML-2 in the appropriate phylogenies (data not shown). With these elements, and those discussed below, it is likely the HML-2 elements presented in this study are an exhaustive and most current representation of the available sequenced genome builds and referenced literature.

It was a concern that at least four of the most recent HML-2 proviruses were not present in the current (and past) human genome builds. At least three were referenced from the literature and from GenBank. The K105 solo LTR was putatively located as a on an unaligned chromosomal region by virtue of published flanking regions, however through comprehensive searching, a BAC was identified whose full-length sequence

BLATed to the chimpanzee K105 locus (builds panTro2 and 3). The respective flanking sequence also hit with best score to the U219 contig in Hg19, supporting its identification. The 10p12.1 provirus, K103, was sequenced previously and from GenBank, as was the 19p12b element, or K113. We have also determined the full-length sequence from a polymorphic HML-2 integration, namely the 12q13.2 provirus, presented here for the first time. The lack of these proviruses in current genome builds, with particular attention to the 12q13.2 provirus, is troubling due to the likelihood that novel polymorphic HML-2 will be described in future studies. Indeed, work from our lab (both past and present) from high-resolution DNA hybridizations have suggested a high level of polymorphic HML-2 exist within humans [89] (J.H.W. and J.M.C., in preparation; also in Chapter 4). As new elements are described, their inclusion (or at least pointed indication) into updated genome builds will help future study of these elements, for example in association to humans disease, and add ease to their verification.

In our preliminary searches and attempts at HML-2 identification, a certain amount of inconsistency and disagreement within the literature was quickly made apparent. Many previous studies of HML-2 in the human genome reference these proviruses by the corresponding BACs, which provide little to no useful information about the proviruses described. Comparing information on individual HML-2 proviruses has been further confounded due to the tendency of inconsistent the nomenclature from different reports. The same proviruses were many times referenced by different accession numbers, multiple accessions, or the reference accession(s) for the original BAC that corresponds to the sequenced genome region. A few referenced accessions lead to BACs that include more than one provirus [64, 94], and some had no listing within NCBI. The

development of a nomenclature system and its standard utilization in the field would add clarity to studies involving the HML-2 (as well as other HERVs), and would decrease complications in future studies and with discoveries of new loci.

The group of proviruses comprising 4q13.2 and 8p11.1 elements are worth mention due to their classification to subgroups according to RepBase. Within RepeatMasker of the UCSC Genome Browser, the proviruses are indicated to have LTRs belonging to both LTR5A and LTR5B. Although it is possible the two sites represent LTR5A/B recombinants, given the support in shared LTR5A motifs within all LTRs, and their phylogenetic grouping as sister taxa within each LTR-based phylogeny (asterisked in Figure 3.5), an alternative explanation is their 'mis'-classification by RepeatMasker. Thus, the systematic classification of subgroups presented here may in fact provide a better indication of the true subgroup-specificity within the HML-2 group.

The LTR5-Hs represent the most recent germline integrations from the HML-2 group of HERVs, and account for all species-specific integrations.

In this study, we have provided evidence that the HML-2 group is comprised of the subgroups LTR5-Hs, A, and B. Support of each subgroup provided in the complementary analyses based on nucleotide sequence, in addition to the phylogenetic relationships within each group (Figures 3.5 and 3.6). These data indicate the LTR5B elements are ancestral to both the –Hs and A, and the general estimations for times of integration of proviruses within each subgroup support this finding (Table 3.1 and Figure 3.8).

Of the total HML-2 that cold be classified by LTR subgroup, more than half (45, or ~53%) were found to belong to the LTR5-Hs, indicating a relatively high level of germline activity for this particular group. The LTR5A and B were equally divided in number among the HML-2, respectively with 20 and 24 elements, suggesting their lower success of germline colonization leading to humans. Moreover, we provide evidence for the latter two subgroups to have undergone a certain level of genome expansion as a result of segmental duplications in the host, which has contributed to nearly a third of the elements within each subgroup. From studies by others [64, 67, 89] and presented here, it appears that most, if not all, the LTR5-Hs proviruses have originated from exogenous activity, and in this context the predominance of the LTR5-Hs in humans is even more conspicuous.

From the analysis of subtype distribution within the LTR subgroups, we also observed that the type 1 elements appear represented only within the LT5-Hs proviruses, from the total number that could be conclusively typed. To be sure, ~16 elements were for which internal deletions or truncations had removed the type 1-specific feature. Within the remaining LTR5-Hs members, comparison of subtype to the human-specific elements indicated a slightly higher frequency of type 1, however when comparing subtypes among the proviruses that have also been described as polymorphic, the opposite was observed, and type 2 were observed at a higher frequency. Given the relatively small number of analyzed proviruses it is difficult to draw any conclusions; however these observations suggest a trend in which type 2 elements are more frequently observed the most recent HML-2 integrations, whereas the opposite is observed when

compared to the LTR5-Hs group as a whole. In general, there is a tendency in which type 2 elements are more commonly observed among the most recent integrations.

Of the typeable elements, the node corresponding to their appearance is indicated in the LTR5-Hs tree in Figure 3.6 A, however we consider the node tentatively, as a portion of the elements were excluded from the analysis. That said, many of the elements surrounding the node have been analyzed by PCR to determine their species distribution previously in our lab [64, 65]. For the clade containing the type1 proviruses, the oldest to be detected was the provirus at 3p25.3, a recombinant provirus, detected in the orangutan. Elements that were formed prior to the type 1 node, for example 11q12.3 and 19q13.12b were detected in the gibbon. With the caveat that PCR analysis for the 5p12 element is lacking, we can infer the putative type 1 emergence was at least around the time of the gibbon/Great ape split, or ~15-18 mya [69]. Following its appearance within the LTR5-Hs elements, type 1 is the only observed subtype until the formation of human-specific proviruses [94, 115]. The type 1 proviruses are not monophyletic within the clade, and following the appearance of the human-specific elements, are dispersed with type 2. The type 1 Δ292 bp evidently arose from a single origin, as phylogenetic analysis of the flanking regions of both subtypes is said to demonstrate the monophyletic branching of the type 1 [93].

Possible explanations for the higher frequency of type 1 within the LTR5-Hs are the removal of type 2 elements over time in other species, however this fails to explain the eventual success of type 2 within the human species. As the type 1 share the same Δ292 bp, another explanation for this pattern is their carriage by functional proteins from intact elements, in a mechanism of complementation *in trans*, although this

mechanism of proliferation is relatively rare among HERVs, with the exceptions of the HERV-H, ERV9, and a subset of the HERV-W groups [71]. As a result of the mode of copying, and in contrast to the HML-2, the elements belonging to those classes have been able to reach much higher copy numbers, whereas the HML-2 have remained comparatively low in total number. A prediction from *trans*-complementation would be a higher d_N/d_S for the type 1 *env* than the type 2 HML-2. One analysis has suggested the d_N/d_S does not differ significantly different between the two subtypes, interpreted as evidence for type 1 elements to have been under purifying selection [93]. If this were true, it would imply an even greater success of the type 1 elements in the recent primates, although the potential infectious capacity of such elements carrying the 292 bp deletion is unknown. An alternative explanation –and one that seems to be more accepted in the field– is the re-generation of the type 1 allele through recombination or gene conversion of the relatively small Δ292 bp region multiple times in the germline [93, 186].

Some LTR5A and B proviruses have experienced copy number increases as the result of host chromosome segmental duplications.

Based upon our estimates, the LTR5A and B elements entered the genomes of primates, on average, ~20-30 mya, approximately at the time of the split of the apes from Old World monkeys (~25 mya [189]). However, the success of each subgroup to increase in copy number within the germline has not been solely through exogenous activity, particularly for the LTR5A and B, as a third of each subgroup can be attributed to amplification via segmental duplications of the host genome. In contrast is the LTR5-Hs,

which lack evidence of having been duplicated, save the tandem duplication of the K108 provirus at 7p22.1 detected in some individuals.

At least 10 LTR5A were represented as duplicates within the genome, including 7 proviruses and 3 solo LTRs. By close examination of the surrounding host genome surrounding each site, we verified a pattern of flanking REs that were shared between most of the LTR5A sites. The host regions that differed were observed to have little variation (as inferred from Figure 3.12 C) from what we deemed the overall 'core' structural order. The shared sequence was determined to be in excess of 200kb in all cases, providing a strong indication that the regions were the result of large-scale segmental duplications. Parallel literature searches led to the identification of two reports of the large 'composite-like' segmental duplicon, DA, which has contributed to at least 20 segments of >200kb in humans. Close comparison of the DA 'core' region and the core motifs identified in the context of each LTR5A copies confirmed the structural order and patterns of core REs agreed, providing an indication that the LTR5A were associated to these specific segmental duplications.

The DA segments are non-randomly distributed within the genome and are found predominantly within or near chromosomal regions involved in large inversions, duplications, and evolutionary breakpoints [193, 194]. These segments were originally noted in 2003 [195] and later detailed in the 2009 study by Ji, *et al*., and were reported in 20 copies distributed non-randomly throughout the genome, at specific chromosomal sites analogous to most of the LTR5A copies, with the exception of their identification of at least one segmental duplication on chromosome 7, within which we observed no HML-2 elements. We were able to confidently map each of the LTR5A proviruses to

their respective duplicated segments, and as expected, each could be mapped over the same relative site (Figure 3.12 and Appendix A). As a reference, the 8p23.1c and d copies were used, as observed from the same snapshot within the UCSC Genome Browser. Additionally, by BLAT of the flanking sequence surrounding each LTR5A duplicate, an additional 3 segments were identified with solo LTRs at the respective copy site, indicating an even more robust copy increase from this particular cluster in the absence of infection. Further support was inferred from their phylogeny, in which the solo LTRs group within the branches of the LTR5A duplicates with the common outgroup of the 3p12.3 element (Figure 3.11). Together, these observations indicate this specific segmental duplicon, 'DA' has contributed to the copy number of the LTR5A.

With reference to the phylogeny of the LTR5A duplicated elements, the estimated ages of the proviruses at the site, and with the inferred information of the reported species distribution of the DA segments, a putative model can be traced for the duplication-mediated expansion of these particular LTR5A within the genome (Figure 3.12 D). At least two explanations can be offered for the presence of the LTR5A copy site in these specific types of chromosomal segments. (1) The LTR5A copy has been associated to a DA segment from which subsequent duplicates evolved, in a stepwise manner as reported for the DA, gaining first the HERV-E insertion into the core RE, and later the tandem arrangement and minor duplication from 8p (indicated in Figure 3.12 A). Such a model would require duplication of at least two lineages of DA in order to account for copy number. (2) The LTR5A copy site has been exchanged between DA segments in gene conversions, and given the branching pattern of their LTRs, conversion of the full-length copy is most likely.

Given the large segments of shared sequence, homogenization by gene conversion is difficult to conclusively dismiss, and could lead to underestimations of age and remove the divergence patterns. From the analysis of a subset of duplications in primates, at least one study has reported consistency between species divergence and nucleotide substitution rates, arguing against predominant affects of gene conversion on duplications [196]. Any variation between genome regions and/or the particular segments involved remains unclear, however our results suggest (at least from the available human sequence) an absence in this particular set of HML-2-associated duplicons. Direct species comparisons should help to clarify this issue in the future.

Quite interestingly, at least three additional HML-2 proviruses were found in association with DA segments. The UCSC Browser windows provided in Appendix A gives a clear representation of the positions of these sites with respect to the boundaries of each duplicated segment. As briefly mentioned above, the location of the 3p12.3 LTR5A element was within 90kb of the LTR5A site, just on the opposite side of the core REs and clearly within the DA boundary. Similar associations to DA segments were for proviruses belonging to the LTR5-Hs, each within a segment that contained the 'empty' LTR5A copy site, for example, the 3q21.2 K(I) provirus was located within just 5kb of the corresponding LTR5A site The remaining LTR5-Hs site was confirmed to represent 8p23.1a, or the polymorphic K115. The K115 provirus edges a ~100kb region flanked on either side by what appear to an inverted duplication of at least two DA segments; the entire region contains the 8p23.1b HML-2 site and spans over 1Mb (also detailed in Appendix A, '8p23.1b' window). Its insertion site lies proximal to an unaligned gap, and as a result the relative distance to the nearest LTR5A copy site could not be determined

for comparison. However, a recognizable DA core is situated within ~300kb upstream (near the 7Mb within the 8p23.1b window), and likely represents the closest corresponding LTR5A site.

The 8p23.1 region warrants further discussion for a few reasons as detailed here. Together, the duplicon blocks that respectively include the 8p23.1a. and b, and the 8p23.1 c and d HML-2 flank a chromosomal region that extends 3.8-4.5 Mb in size and contains at least 50 genes (referenced to RefSeq). Inversion of this entire region was first described in the early 2000's, detected by FISH analysis at population frequencies around ~26-27% in tested Europeans and Japanese [197, 198], and later in up to ~60% of screened individuals in some populations of European descent [192]. These observations were in contrast to the contemporary genome build (and current genome build, personal observation), suggesting the inversion is actually the most common orientation in at least some populations, and that the published orientation represents the minor variant [199]. Heterozygotes also appear in relatively high frequencies (up to 50% of subjects in one study) [200], providing further support for the prevalence of the inversion.

Each of the segments is associated with cellular genes in two contexts: those between the inversion region (flanked by duplicons), and those within the segments themselves. A report in 2009 demonstrated that the expression patterns of at least four internal genes were significantly altered depending on the orientation of the 8p23.1 locus; examples of both up-and-down-regulation were observed, and in each case the expression pattern was such that heterozygotes (inversion/non-inversion) were of intermediate phenotype [200]. The cause of increased expression was unclear, but the authors did posit that the duplicons contain "regulatory variants" affecting gene transcription within the

region. The 8p23.1a/b and c/d segments themselves contain genes, some of which are present in multiple copies. The most well-characterized are probably the *beta-defensins*, which are found as copies of a single ancestral cluster as a result of segmental duplication within the 8p23.1 locus (RefSeq summary NM_001037668). Interestingly, within the 8p23.1a/b and c/d segments, each of the associated HML-2 proviruses lies within ~10kb of an exon, and it is possible the presence HML-2 may influence the expression of the genes. Also, at least one site, K115, is polymorphic within humans, and could have further unknown affects on the transcription from this region in some individuals. However, the duplicons themselves are concentrated in REs, including other HML-2 and HERVs, that might also affect transcription within the region.

Also in this context, the large duplicon is present at several other sites within the genome in a non-random distribution. Of the ~20 DA segments that were documented, the larger complex loci at 8p23.1 appear to have arisen most recently, as suggested from conservation in nucleotide sequence and ordered structure, and from analysis of species distribution as detected by others [193, 194]. The 4p16.1a segment is highly similar, and the apparent ancestral copy to those at 8p23.1, and also within a few kb of several host genes, though the linkage is not as defined as on chromosome 8 (also refer to the 4p16.1a/b window in Appendix A). Again, the exact significance of these observations is unclear, one may speculate that there is a certain propensity for ERV integration (or other REs) within these particular genome regions. The segments each contain a number of ERV types, with multiple copies of ERV1, HERV-E, HERV-H, and LTRs from several HERV groups. Also, the more recently evolved type III DA contain additional sets of tandemly arrayed solo LTRs (~10-20 per set are visible) belonging to

the LTR5A subgroup. Thus, these regions appear to be relatively high in ERVs, in addition to other multiples of REs. Although it is not clear that the DA have contributed to human-specific duplication(s), prospectively such an event could lead to the similar formation of duplicates from other HML-2 subgroups.

Finally, the LTR5B subgroup also showed evidence of segmental duplication, however to a lesser extent than those sites within the LTR5A discussed above. The two clusters giving rise to the Xq28 and 1p36.21 elements each had aberrant grouping of LTRs, identical and intact TSDs, and shared flanking genome regions. Interestingly, the overall pattern of duplication appeared highly similar, although there was no detectable larger chromosomal regions as with the LTR5A, and the best hits for the flanking host DNA for the Xq28 and 1p26.21 clusters were the proviruses themselves. Thus, each cluster appears to have arisen through an independent segmental duplication ranging in size from about 50kb, as was the case for the Xq28 copies, to nearly 200kb, as observed for the 1p36.21 duplications.

An interesting trend between these two chromosomal loci is that in each case, a single provirus from each cluster appears to be situated such that it marks the boundary of the segment. Perhaps coincidental, the border elements, 1p36,21a and Xq28a, are also truncated in a similar manner of their 3' halves, and the duplication stems from the resulting edge, through the 5' LTR of each element (a detailed view of each is provided in Appendix A). Also, the truncations result in a size difference of just 118 nt, with 1p36.21a having the added length. The significance of this observation, if any, is unclear. Possibly, the segments corresponding to each cluster were generated from the same mechanism. However, in the absence of mechanistic details, we can only speculate the

pattern observed for the LTR5B clusters, though these elements may have been points of nucleation for the recombination leading to the observed duplication/inversion patterns. As is the case for the LTR5A cluster, there are several directions to pursue in order to further characterize the duplicated elements within the LTR5B, and future analyses should shed some insight into these observations.

Finally, briefly mentioned above were the additional three elements within the LTR5B, located on chromosomes 15 and Y (Figure 3.3 C). showed evidence of having been duplicated from their phylogenetic comparison of sequences corresponding to *pol* and *env* with other HML-2 (data not shown). Each of these elements was identified in our BLAT searches of the K113 *pol* and *env* sequences, and neither was previously described. The elements each appear to have been duplicated following the truncation of either end, in addition to a number of shorter internal deletions (Figure 3.3 C). This suggests the ancestral element was stably integrated for some time prior to duplication, and suggests this event was relatively recent. As preliminary support, their pairwise comparisons reveal a level of nucleotide identity of >99.93% for Yq11.23a and b, and >92.89% between 15q25.2 and either Yq11.23 copy. An extension of these observations would be interesting to further elucidate each of these copies, as they apparently represent some of the oldest HML-2 copies, with respect to their overall nucleotide divergence from K113 and other closely related HML-2, and are the only detected elements of internal reading frames along with evidence of amplification via segmental duplication.

As has been demonstrated previously from work in our lab, endogenous retroviruses provide a useful tool for tracing evolutionary histories within the genomes of primates, for example as a source for estimating primate phylogenies [63], and for the

detection of genome rearrangements from ectopic recombination and from gene conversion [64, 65]. Here we provide evidence that ERVs, and in particular here the HML-2 proviruses, may also serve as indicators of large scale rearrangements from chromosomal segmental duplications, and perhaps even have even had direct involvement in such events.

Previous studies of primates' genomes have indicated a non-random distribution of segmental duplications, in chromosomal context and with reference to genic regions, having occurred in recent evolution [201, 202]. In most recent primate evolution including the great apes, the tendency has been toward intrachromosomal duplications (for example at the 8p23.1 and 4p26.1b and 3b loci), as opposed to events between chromosomes, suggested to have peaked following the divergence from Old World monkeys [196, 201, 202]. Duplications of genome sequence, and especially those events leading to additional copies of genic regions (for example, exons, partial genes, full genes, or pseudogenes), are potential template sources for the evolution of exaptation of functions from existing sequence [203]. In this sense, the adaptation of an existing function has the potential for 'new' genes to evolve relatively quickly, in comparison to the probability for new functions to arise through gene mutation alone. Given the distribution we have observed the HML-2 associated with such duplications, in particular with segments corresponding to the so-called DA, a role for the recently active HML-2 in such events is possible, with implications in future divergence and speciation.

Finally, it is noted that the analyses presented here are from recent work, and have resulted in at least the equivalent number of potential directions. It is possible other HML-2, or even those identified in future research, are also associated with duplicated

regions; other studies from this lab have provided evidence that some HML-2 have been part of genome rearrangements in the past [64, 65]. A further association with large-scale duplications again postulates the elements in genome rearrangements, but also suggests a possible involvement in the evolution of new coding and/or regulatory functions.

Chapter 4

Results and Discussion Part 2:

The Distribution of Polymorphic HML-2 in Human Health and Disease

Significance

Numerous reports have demonstrated the up-regulated expression of proviruses from the HML-2 group in the tissues associated with several types of cancer, neurodegenerative disorders, autoimmune diseases, and elsewhere. Analyses of bulk cDNAs from a few such tissues has shown that the majority of HML-2 transcripts originate from the most conserved proviruses [144, 151, 158]. Transcripts from at least 23 proviruses have been detected, including all but 6 human-specific elements, and from which most polymorphic proviruses are observed depending on the sample set [151].

The overall expression of HML-2 proviruses follows a pattern of tissue specificity, as transcripts and proteins have been detected in tumors but not adjacent epithelia . Furthermore, the proviruses themselves exhibit different levels of expression in different tissue types. For example, transcripts corresponding to the 1q22 (K102) provirus were shown to be represented in nearly >70% of cDNAs analyzed from breast tumor tissues [151], most transcripts from germ cell tumors have been from 22q11.21 (K101) [147, 151], and the most frequently detected transcripts in the brain were from 5q33.3 (K10) and 3q21.2, or HERV-K(I) [151]. It is unclear whether members of the HML-2 group exert pathogenic effects, or are expressed as a consequence of deregulated expression in diseased tissues.

In animals, ERVs related to the HML-2 group have been well-studied for their pathogenic effects. Two established examples have been from MMTV and MLV, which cause mammary tumors and leukemia in mice, respectively, from the deregulated expression of proto-oncogenes within the host in following novel provirus formation in a mechanism of insertional mutagenesis [12]. Another *Betaretrovirus*, JSRV, induces a

contagious type of lung cancer in sheep by oncogenic properties of its Env protein [142]. It has been suggested that HERVs may have the capacity for similar affects. In humans, most HERVs are fixed and heavily mutated, having been from ancient germline infections, and it seems unlikely that such elements would have a role in a condition that affected a fraction of the population, on the basis that a provirus with pathogenic effects would likely encode functional proteins. Given the recent replicative activity of some members of the HML-2 group, there has been interest in the polymorphic proviruses for potential involvement in several such diseases. A relationship between HERVs and human disease may be suggested by the demonstration of a genetic association of an inherited provirus with individuals at high disease risk. To this end, our goal has been to analyze the genetic association between polymorphic HERVs and disease.

Of the described polymorphic HML-2, the K113 and K115 proviruses have been given particular attention for association to disease [96, 167, 204, 205]. Present respectively within ~30% and ~15% of individuals tested, K113 and K115 are estimated to have integrated into the germline <2 mya and have retained functional ORFs [95]. In particular, the full-length K113 provirus is intact, and has even been shown to express particles [135, 136]. Also, the rate of germline formation appears to have been constant since the divergence from chimpanzee, suggesting an active 'pool' of infectious proviruses exists within humans [93]. However, no naturally occurring infectious element has been identified.

A few reports have provided evidence that additional, as yet uncharacterized, insertional polymorphisms of HML-2 elements are detectable within humans [89, 131]. One of the most compelling lines of evidence was from work in our lab in 2004, in which

a panel of human genomic DNAs was analyzed for HERV-K (HML-2)-specific elements by unblot, revealing a surprising amount of detectably polymorphic elements [89]. A comparative PCR analysis lent support for the identification of about 6 of the proviruses by shared banding patterns, however examination of the blot revealed a few additional bands of varying frequency within the sample set.

Aside from K113 and K115, no other polymorphic HML-2 has been investigated with respect to disease. Such proviruses would likely have formed recently, with low frequencies within the population, and would be predicted to have retained some functions with possible replicative capacity. However, the lack of information for all HML-2 proviruses makes the inference of disease association one of difficulty. In order to examine the possibility of a novel insertion having prevalence in human disease, we used high-resolution unblotting to analyzed the genomic DNAs from patients of two conditions with implicated involvement of HML-2, breast cancer and schizophrenia, and individuals without a history of the disease.

Case-control analysis of described polymorphic HML-2 proviruses in breast cancer

Investigation of polymorphic HML-2 in a subset of breast cancer patients

Previous work from our lab and by others has led to the identification of about 10 examples of proviruses belonging to the HML-2 group for which multiple alleles were detected with varying frequencies among humans (Table 4.1) [89, 90, 95, 173]. Additionally, a search of the available literature led to the 2007 report by Belshaw, *et al*. of an uncharacterized HML-2 polymorphic integration located at 12q13.2, for which all three alleles were detected in PCR screens of human genomic DNAs (asterisked in Table 4.1) [88]. As described in the results from analyses of the HML-2 group in Chapter 3, we verified the polymorphic status of the 12q13.2 site by PCR screen of human genomic DNAs, and sequenced and characterized the full-length provirus (Figures 3.1 and 3.2) (R.P.S., J.H.W., *et al*., in preparation). There were no other reports of HML-2 loci with multiple alleles that included proviruses, and to the best of our knowledge, there have been no such sites identified since. Of the 11 identified loci, at least 5 elements are not yet fixed within humans, as indicated by the presence of the pre-integration site in a portion of individuals tested, whereas the corresponding empty site has not been detected at any of the remaining 6 loci.

We verified the chromosomal locations for 9 of the 11 polymorphic proviruses within the Hg19 genome build in parallel BLAT-searches against earlier genome builds (Mar.2006 Hg18; May 2004 Hg17; July 2003 Hg16). The full-length sequences of the remaining two elements, namely K113 and the provirus detected at the 12q13.2 locus, are not represented within the published sequence builds, however 12q13.2 is represented as the solo LTR allele. The flanking region of either provirus is

Table 4.1. Polymorphic HML-2 proviruses described in human DNA.

HERV-K notation	locus	start (bp)	Sub-type[b]	Detected alleles[c]	Accession number	Reference
	1p31.1	75842771	1	pro	AC093156.2	[89]
K106	3q13.2	112743479	1	pro, solo	AC024108.22	[173]
K109	6q14.2	78427019	2	pro, solo	AC164615.1	[173]
K108[a]	7p22.1[a]	4622057	2	pro, solo, tandem, pre	AC164614.1	[90]
K115	8p23.1a	8054700	2	pro, pre	AY037929.1	[95]
K103	10p12.1	27182399	1	pro, solo	AF164611.1	[173]
	11q22.1	101565794	2	pro, solo, pre	AP000776.5	[186]
	*12q13.2	55727215	1	pro, solo, pre	na	[131]; this report
	12q14.1	58721242	2	pro, solo	AC074261.3	[186]
K113	19p12b	21841536	2	pro, pre	AY037928.1	[95]

[a] Published as a tandem provirus as the product of an allelic recombination. The start coordinate provided above refers to the K108-R, or right, provirus of the tandem pair.

[b] Subtype is with reference to the 292 bp deletion diagnostic of type 1 proviruses.

[c] Pro, provirus; solo, solo LTR; pre, pre-integration (empty) site.

publicly available [88, 95], and BLAT searches were performed to verify the integration sites. With reference to the Hg19 genome build, we designed primers that would amplify the 5' LTR of each element, to infer the presence of the provirus at that site. Additional primer sets were designed within the flanking region of each provirus to detect the presence of the remaining described alleles per site.

We analyzed the frequencies of each polymorphic integration in a case-control screen in order to provide an indication of the prospective observed HML-2 detection during DNA subsequent unblotting, but also with the expectation that focus would be given to individual proviruses should there be any strong prevalence observed between groups. For these purposes, we screened a panel of DNAs from diagnosed breast cancer patients and from individuals with no history of the disease. All samples were generously provided from the ACS and were from the CPSII Nutrition Cohort [179]. The CPSII collection is from a large-scale study designed to provide a means with which to

investigate the relationship between lifestyle factors and exposure risk to cancer incidence, mortality, and survival. The cohort, represented by more than 84,000 men and 97,000 women, has been tracked over the last twenty years to gather information concerning reported cancer incidences and causes of death, each of which has been individually verified.

From the CPSII DNA collection, we were provided with an initial set of 100 blinded samples for examination: 50 were from breast cancer cases or controls (n=25 per group), and 50 were represented equally between prostate cancer cases and controls. All initial analyses were performed while the sample sets were blinded. Following release of the keycode, samples were sorted according to disease group, and all analyses repeated for the breast cancer groups in order to verify initial results and for a direct case-control comparison.

In preliminary screens of the blinded sample set, two PCRs were used to confirm the presence or absence of alleles present at each polymorphic site using primers spanning the 5' LTR of each provirus (indicating the presence of the provirus), or with primers spanning the integration site (to detect either a solo LTR or pre-integration sequence) (also refer to Figure 3.1). Representative products from each amplification were sequenced in both directions to confirm the correct product and to rule out non-specific amplification (data not shown). Upon completion of initial PCR screens, the keycode was released and the samples unblinded and sorted by group. With DNA samples grouped by case-control identity, PCRs to detect the presence of the provirus were repeated for each polymorphic site to confirm the initial results and for a direct comparison per group. The frequencies of each provirus were calculated for either group

and statistically analyzed by χ^2, with a *p*-value of <0.05 regarded as significant within the dataset. The results are summed below.

Table 4.2. Prevalence of polymorphic HML-2 in breast cancer.

HML-2 locus	Breast cancer cases[a]		Healthy controls[a]			
	# positive	frequency	# positive	frequency	χ^2	*p*-value[c]
1p31.1	16	0.64	17	0.68	0.09	0.76
3q13.2	25	1.0	25	1.0	undef.[b]	
6q14.2	21	0.84	23	0.92	0.75	0.34
7p22.1R	25	1.0	25	1.0	undef.	
7p22.1L	24	0.96	25	1.0	1.02	0.31
8p23.1a	6	024	1	0.08	4.15	0.04*
10p12.1	24	0.96	25	1.0	1.02	0.31
11q22.1	23	0.92	20	0.80	1.49	0.22
12q13.2	20	0.80	21	0.84	0.13	0.72
12q14.1	23	0.92	22	0.88	0.22	0.67
19p12b	3	0.12	3	0.12	undef.	

[a] Total sample size was 50, split equally over case and control samples (n=25 per group).

[b] Chi-square value was undefined as a result of equal frequencies observed between cases and controls.

[c] All statistical analyses were performed by the Data Design and Resource Center at Tufts University; *p*-values were calculated based on a total sample size of 50.

For most polymorphic integration sites analyzed, we found no association of a particular provirus with either sample group within this sample set. Unexpectedly, we did observe the K115 provirus to have a higher prevalence within the cases (6/25, or to a frequency of 0.24) than in control group (1/25, or 0.04), with a *p*-value of 0.04. As mentioned, this particular provirus has been analyzed in the context of a few human diseases (including the genomes of breast cancer patients [205]) for evidence of genetic association, however no significant difference has been substantiated. To be sure, we

tested the significance of this result in a larger screen. Although this *p*-value falls within significance for the given dataset, our strategy for the analysis of each site involved more than one comparison between the two groups. Due to the increased number of comparison performed, such a difference was given an increased likelihood to be observed among all comparisons than in the case that a single comparison was made. Thus, we sought a higher level of support in order to strengthen the probability of a single outcome given the data set.

To accommodate for testing with multiple comparisons, we analyzed the significance of the observed prevalence of K115 in a larger-scale screen. From the initial analysis, we observed a 16% difference in the case-control frequencies for the detection of the K115 provirus. A power analysis was performed with the help of the Design Data and Resource Center (DDRC) at Tufts University to determine the minimal number of samples to replicate the same difference in case-control frequencies to a statistically significant level. This type of analysis was also appropriate to compensate for a smaller sample size in the initial screen in order to strengthen the statistical power in the analysis.

We were provided an additional 200 samples (100 per cases or controls) from the CPSII cohort, which allowed for testing to a significance of 0.05 with greater than 90% power. All additional samples were screened while blinded and those samples positive for K115 confirmed in a second PCR. Subsequently, we were provided the case/control key for those positive samples alone. From the analysis of the larger sample set, we observed K115 to be present in 6/100 cases (0.06) and 11/100 controls (or 0.11) (corresponding to a χ^2 of 1.61 and *p*-value of 0.20). While we cannot rule out the possibility that K115 may be implicated in a subset of breast cancers, these results

suggest the described polymorphic HML-2 proviruses are not associated with an increased disease risk, at least for the sample set.

Most of the described polymorphic HML-2 have been detected at relatively high frequencies within humans; even the K113 and 115 proviruses have been observed in up to 30% to 40% of tested individuals, with dependence on the ethnicities of the sample tested [95, 96]. Also, the vast majority of HERVs are fixed within humans and are heavily mutated, making it difficult to hypothesize such elements could be associated to a disease that affects a fraction of the population. Proviruses that are not fixed within humans have been considered a more appropriate source for potential involvement in disease, given the increased likelihood that such elements encode intact and functional proteins, and their presence in a subset of the population. Also, polymorphic elements with sub-deleterious effects may persist in low allele frequencies. As the only HERVs with such recent germline activity, the HML-2 group contains representatives of the most likely candidates for such a scenario. Aside from the 11 described polymorphic integration sites, it has been suggested that additional polymorphic HML-2 are present within humans [76, 89, 93, 131]. Previous work in our lab has shown that polymorphic elements are detectable in the high-resolution DNA hybridization technique referred to as 'unblotting' [89, 180, 206]. We used this approach to estimate the total number of polymorphic HML-2 within our sample set, and to infer whether an undescribed element might have a higher detectable frequency in either group.

***In silico* analysis of polymorphic HML-2**

An *in silico* restriction analysis was performed in order to identify appropriate restriction enzymes for subsequent DNA blots to visualize HML-2 proviruses, and to generate predicted fragment patterns of HML-2 with reference to the published human genome build. In a similar strategy to that described above (also refer to Results and Analysis in Chapter 3), we BLAT-searched the Hg19 genome build to identify proviruses with high percent identity to K113. Briefly, for a total of 62 elements, full-length sequences were extracted and used to generate an alignment, to which 4 other described proviruses were added, located respectively at 10p12.1, 19p12b, an unaligned contig U219 (K105), and 12q13.2 (R.P.S., J.H.W., *et al.*, in preparation).

We searched the alignment for sequence regions which were highly similar among the most recently integrated elements, but distinct from the corresponding regions of the remaining HML-2. Close examination revealed a conserved ~32 bp region within the *gag* leader region of the proviruses most similar to K113. BLAT of this region alone returned identical hits for 17 HML-2 proviruses, and another 8 with two or fewer mismatches (Figure 4.1). Included were all human-specific elements and published polymorphic proviruses, supporting the specificity of the site. This sequence, which we refer to as 'K-seq', was used for further *in silico* restriction fragment analyses and in subsequent unblotting to infer the presence of the most conserved HML-2 proviruses.

Unblotting, or DNA hybridization in dehydrated agarose, is similar to Southern DNA hybridization, but offers the advantage of increased resolution without loss of target DNA during transfer with the caveat that at least 10μg of digested template is required per sample per run [180, 206]. In order to preserve the limited amount of

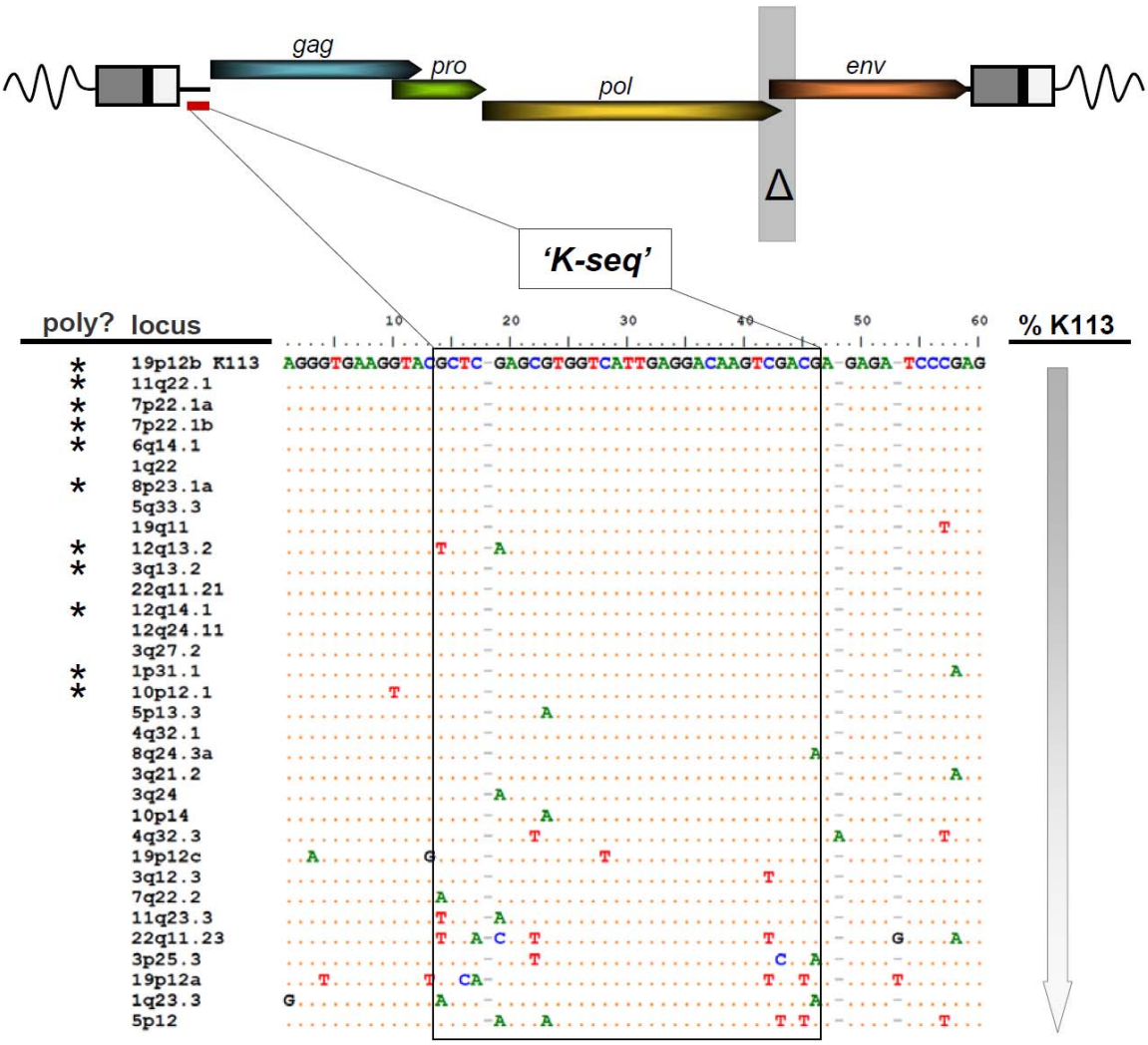


Figure 4.1. Nucleotide sequence alignment of the 'K-seq' oligonucleotide site conserved in the most recent HML-2 proviruses.

A BLAT of the ~32bp K-seq site (shown in box) returns each of the proviruses included in the alignment. There were 17 proviruses with a 100% match to the region, 8 with one base changes with reference to K113, and several additional hits with >2 base changes. Each element has been aligned with reference to K113 to depict the conservation of the nucleotide sequence within the K-seq target site, and the specificity to the most recently formed germline integrations. A dot indicates bases that are shared with K113, and differences are indicated by the base present at that site. The proviruses have been arranged as a function of their percent nucleotide identity to K113, as inferred from individual pairwise comparison of each full-length. The asterisks at left indicate described polymorphic proviruses. Lines above box are used to show the relative position of the K-seq site downstream of the 5'LTR but within the untranslated region upstream of the *gag* coding frame (indicated with a red line).

DNA available per CPSII sample (about 1μg in total), we subjected each sample to a whole genome amplification (WGA; Qiagen), in which Φ29 polymerase catalyzes the branching amplification from random hexamers in a reaction suggested to generate minimal sequence bias. Preliminary unblot hybridizations revealed at least one limitation of using WGA-DNA as a template, such that fragments larger than ~5-6kb in length were not detectable (data not shown). A possible explanation is the presence of a higher concentration of completed double stranded DNA within this size range in each reaction. To compensate, we performed an *in silico* restriction analysis to identify enzymes that generated ~1-6kb fragments representing the junction sites of each conserved HML-2 provirus (Figure 4.2).

We analyzed each provirus for enzymes predicted to cut at least once within the provirus but not within the 5'LTR, with reference to the K-seq site. About 35 enzymes were further analyzed for predicted fragment sizes as inferred from the nearest restriction site in the flanking host regions. In this analysis, each detectable fragment is predicted to contain a single proviral junction site, the size of which is dependent on the distance from the 5'LTR edge to the nearest restriction sites in the host flanking DNA. Of 6 potential restriction enzymes, two were chosen for use in unblotting, *Bsr*I and *Dra*I, based on the distributions and total number of fragments predicted to be within the detectable size range for the analysis of WGA-DNA (Figure 4.2). Plotting the respective fragments logarithmically generated a predicted banding pattern for comparison to the subsequent unblot analysis.

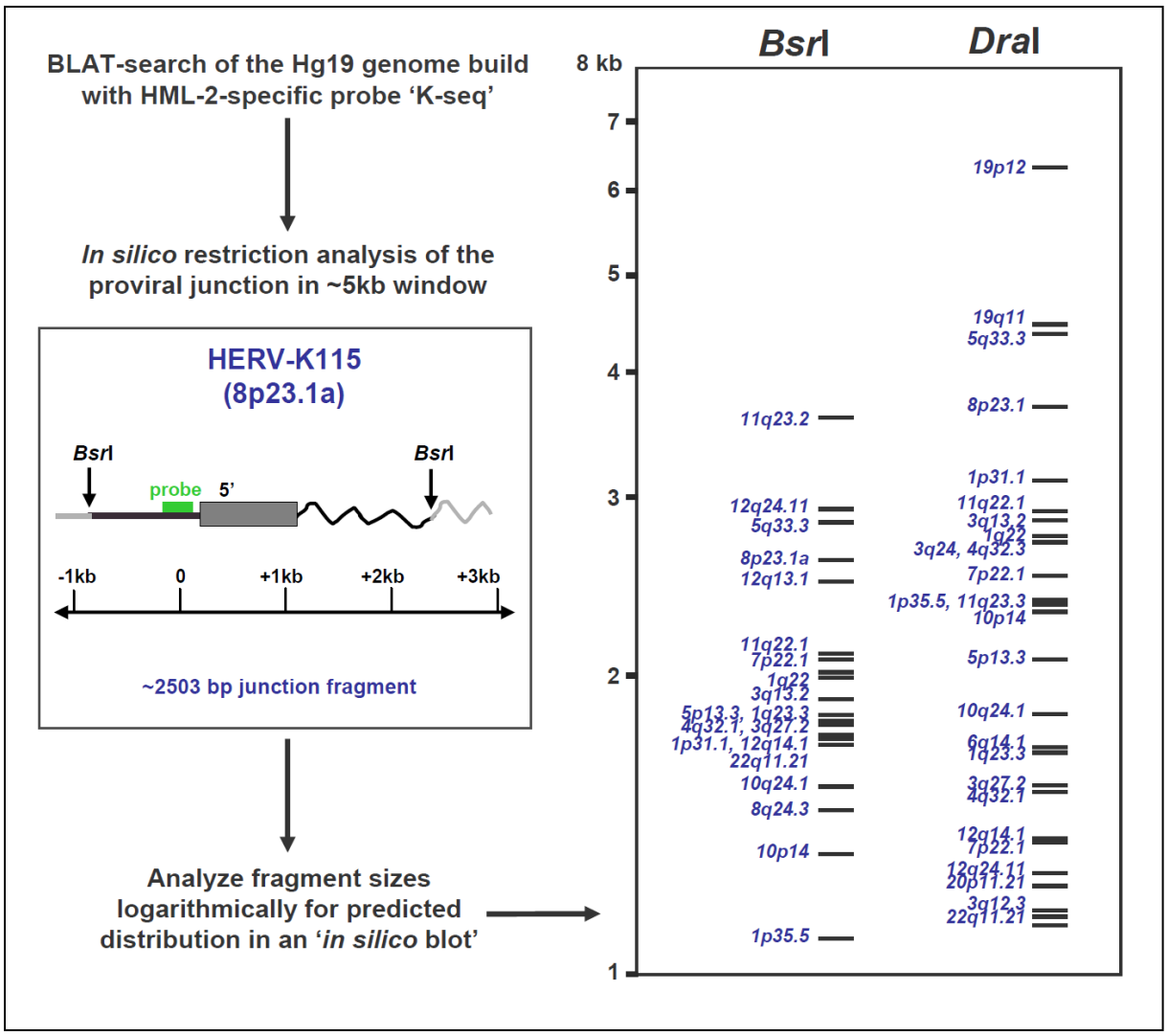


Figure 4.2. Identification of restriction enzymes for use in unblotting.
Shown is a flow chart of the overall strategy to identify useful restriction enzymes for unblotting of WGA-DNA samples. All proviruses identified in BLAT searches of the 32bp K-seq nucleotide sequence were analyzed for restriction fragment patterns to identify restriction enzymes that cut once within the provirus (but not upstream of the K-seq target site, nor within the 5'LTR) and in the host flanking sequence proximal to the 5' LTR. K115 is shown as an example in the figure (drawn to approximate scale). Following tentative identification, individual proviruses were analyzed in an '*in silico* blot' by plotting the predicted sizes of digested fragments that contained 5' junctions of each provirus, inferred by the presence of the K-seq site. *Bsr*I and *Dra*I predicted fragment patterns are shown at right.

Case-control detection of polymorphic HML-2 in breast cancer

We used unblotting [180, 206] of WGA-DNA to infer the distribution of polymorphic HML-2 proviruses within the genomes of the CPSII subjects. As with the PCR detection for previously described polymorphic elements, all preliminary unblot analyses were performed while samples were blinded, and the analysis was repeated as a direct case-control comparison following release of the group identities. Preliminary unblotting was performed in separate runs with *Bsr*I and *Dra*I, and the subsequent case-control comparisons were analyzed with *Bsr*I alone. Genomic DNA from the T47D breast tumor-derived cell line was also run as an added control and for the comparison to WGA-DNA. Briefly, WGA-DNAs were individually digested with the appropriate restriction enzyme, and the products separated by electrophoresis through agarose. Subsequently, the gel was dehydrated, and the fragmented WGA-DNAs were hybridized with a radio-labeled oligonucleotide complementary to the K-seq site while immobilized within the agarose gel. The gel was then exposed to film in order to visualize the distributions of detected fragments. The resulting unblots and corresponding *in silico* predictions are shown in Figure 4.3.

As with the *in silico* characterizations, the unblotted samples were interpreted such that each detected fragment represented a single proviral junction site, the size of which was dependent on the length of host sequence to the nearest flanking restriction site. Excluding those junction sites shared among all samples, several visible fragments of similar sizes were observed among both groups, and varied in frequency from about ~0.02 to 0.98. For example, in lane 25 of the control group at ~1400 bp is a single band not observed in any other sample, whereas the opposite is observed for the band visible

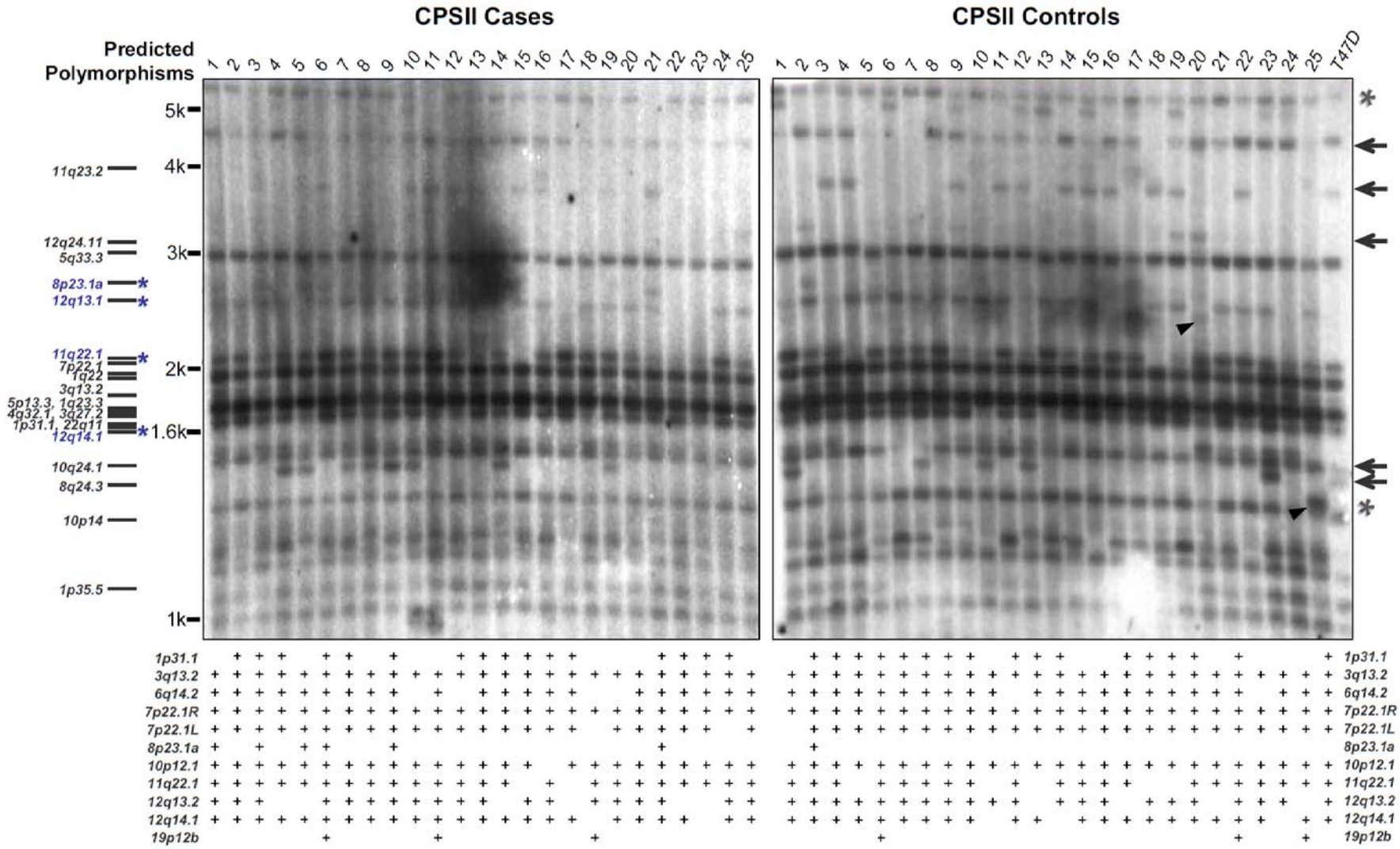


Figure 4.3. The case-control distribution of polymorphic HML-2 proviruses in breast cancer. Samples were unblotted according to group with a HML-2-specific oligonucleotide to visualize polymorphic HML-2. At left are predicted sizes for HML-2 containing fragments. Results from PCR analysis of described polymorphic proviruses are shown below for comparison; a plus indicates detection of the tested provirus. In blue and asterisked are the confirmed polymorphic elements. Fragments observed with variable distribution, but could not be inferred by comparison to PCR analysis or in silico predictions have been indicated at right with arrows.

around ~5500 bp that is present in the majority of DNAs (each is indicated by asterisk near relative fragment size in Figure 4.3). On average, we observed between 18 and 22 bands per lane, depending on the sample, and providing some support that the hybridization pattern corresponded to the most conserved HML-2 elements.

Across all samples, we observed at least 10 proviruses present in some samples but not others, excluding a few regions within each unblot which we were unable to analyze due to low resolution. As expected, the direct comparison of the sample distribution of polymorphic HML-2 confirmed by PCR analysis allowed the provisional assignment of a few fragments based on shared banding patterns between each analysis, including 11q22.1, 12q14.1, 12q13.2, and K115 (Figure 4.3, *lower*). These particular proviruses were lent further support by comparison with *in silico* size predictions of those junction sites (Figure 4.3, at left). We were able to conclusively assign two fragments to their respective sites following PCR amplification of the 5' LTR from the unblotted gel as part of a separate project (11q22.1 and 12q13.2; data not shown).

We examined each unblot for variable bands that were not identified by *in silico* analysis, and whose sample distribution was inconsistent with data from PCR screening. Several fragments met these criteria, and were visible across all samples (indicated at right by estimated size in Figure 4.3). We estimated the frequencies for each fragment by visual inspection of each sample according to group, and compared the case-control frequencies of each by χ^2 analysis (Table 4.3). Consistent with PCR analyses described above, the distribution of each observed fragment did not differ significantly between groups. Collectively the results indicate that at least within this sample set, polymorphic HML-2 are not associated with an increased risk of breast cancer.

Table 4.3. Inferred case-control frequencies of polymorphic HML-2 in breast cancer.

Observed band (bp)[a]	Breast cancer cases		Healthy controls		χ^2	*p*-value
	# positive	frequency	# positive	frequency		
4600	25	1.0	22	0.88	3.19	0.07
3700	10	0.4	11	0.44	0.08	0.78
3200	1	0.04	4	0.16	2.00	0.16
1500	25	1.0	23	0.92	2.08	0.15
1470	8	0.32	5	0.20	0.93	0.33

[a] Band sizes are approximate fragment lengths; each is indicated by arrow in Figure 4.3. The fragment sizes and banding patterns were not consistent with PCR and *in silico* data, and were tentatively interpreted as uncharacterized proviruses.

While it is probable that at least some of the variable bands represent recently integrated proviruses which are present in just a portion of individuals, there are a few alternative explanations. For example, variable bands could represent full-length proviruses which have undergone solo LTR formation, or exhibit variable presence due to restriction site polymorphism. Gaining sequence information for additional HML-2 will likely help to clarify these issues. Regardless, given the sample size screened here, it is likely some of the observed bands represent as-of-yet uncharacterized loci. Whether they are expressed with other HML-2 in diseased tissues is still in question, however HML-2 transcripts with no apparent match have been reported (examples include [134, 151] and R.P.S, personal communication; but also see [181]). The results are further discussed.

Summary of polymorphic HML-2 proviruses in Schizophrenia

The upregulation of HML-2 transcriptional activity has also been associated with neurodegenerative disorders, in particular schizophrenia and bipolar disorder [112, 164, 207, 208]. A previous study by Frank *et al.* in 2005 reported significant differences in the transcriptional activities of a subset of HML-2 in a *pol*-based microarray analysis of cDNAs from post-mortem brain tissues of schizophrenics [25, 26], and a 2008 study from Flockerzi *et al.* demonstrated the transcriptional upregulation of at least 16 individual loci within the associated brain tissues of schizophrenics [151]. Under a similar premise to the analysis described above, we examined the prevalence of polymorphic HML-2 in genomic DNAs in the context of this neurodegenerative disorder.

For this analysis, we were generously provided samples from the Stanley Medical Research Institute of 105 genomic DNAs divided equally among clinically diagnosed patients with schizophrenia, bipolar disorder, and undiagnosed controls matched for age, race, and sex (n=35 per group). As for the CPSII sample set above, all samples were provided as blinded and the keycode released following the initial analysis of the total set. Samples were then sorted by group and, and all analyses repeated for the DNAs from diagnosed schizophrenics and matched controls. Of note, we were unable to generate any PCR or unblot data for a single genomic DNA from the control group, despite repeated attempts and optimizations, and this sample was excluded from the analyses, resulting in a control sample size of n=34 for all analyses.

Each sample was screened for the presence of each described polymorphic HML-2 provirus in a direct case-control comparison, and the frequencies per site subjected to a χ^2 analysis (Table 4.4). As expected, we found no association of a particu-

Table 4.4. Prevalence of described polymorphic HML-2 proviruses in schizophrenia.

HML-2 locus	Diagnosed cases[a]		Matched controls[a]			
	# positive	frequency	# positive	frequency	χ^2	*p*-value
1p31.1	28	0.80	25	0.74	1.15	0.28
3q13.2	34	0.97	34	1.0	0.98	0.32
6q14.2	35	1.0	33	0.97	1.04	0.31
7p22.1R	35	1.0	33	0.97	1.04	0.31
7p22.1L	35	1.0	34	1.0	undef.[b]	
8p23.1a	3	0.08	1	0.03	1.00	0.32
10p12.1	35	1.0	34	1.0	undef.	
11q22.1	32	0.91	25	0.74	3.84	0.04*
12q13.2	26	0.76	25	0.74	0.05	0.82
12q14.1	31	0.89	29	0.85	0.16	0.69
19p12b	1	0.03	6	0.17	4.13	0.04*

[a] For cases, n=35; for controls, n=34 (also refer to the text).

lar provirus with either sample group for most sites analyzed. There were two exceptions, the 11q22.1 and 19p12b (K113) proviruses, that were observed in opposite magnitude. The 11q22.1 integration was observed at a higher frequency within the cases group (32/35, or a frequency of 0.91) than in the controls (25/34, or 0.74), with a *p*-value of 0.04. In contrast was K113, which was observed at a lower frequency among the cases group (1/35, or 0.03) than controls (6/34, or 0.17), also with a *p*-value of 0.04. Testing these observations to a statistically significant level of 0.05 to 90% power would require at least 144 samples (divided equally by diagnosed cases and controls). However, we were unable to test these observations to a statistically significant level, despite efforts to secure the necessary DNA samples. Although such results could possibly have significance, without a higher level of support these observed results must be considered with caution.

Using WGA-DNA digested with *Bsr*I, we further analyzed the case-control distribution of detectable polymorphic HML-2 integrations by unblot (Figures 4.4 and 4.5). As above, each unblot was examined for polymorphic HML-2 junction fragments of

unexpected size and sample distribution. We observed several such fragments, some of which were visible across all samples per group. However, low resolution in a few regions of either unblot, especially in respect to the sample group representing DNAs from diagnosed schizophrenics, permitted only a few case-control comparisons (Table 4.5). For the variable fragments at ~1470 and 3700 bp, the differences in case-control frequencies were compared by χ^2 analysis, though neither was found with significance.

Table 4.5. Inferred case-control frequencies of polymorphic HML-2 in schizophrenia.

Observed band (bp)[a]	Diagnosed cases[a]		Matched controls[a]			
	# positive	frequency	# positive	frequency	χ^2	*p*-value
4600	n.d.	-	25	0.74	-	
3700	20	0.57	18	0.53	0.12	0.73
3200	2	0.06	2	0.06	undef.	
1500	n.d.	-	n.d.	-	-	
1470	4	0.12	2	0.06	0.66	0.42

[a] Band sizes are approximate fragment lengths; each is indicated by arrow in Figures 4.4 and 4.5. The fragment sizes and banding patterns were not consistent with PCR and *in silico* data, and were tentatively interpreted as uncharacterized proviruses.

Although we found no evidence for an association of a single HML-2 provirus with disease risk, the unblots revealed a high level of detectable polymorphism from the HML-2 group. In just these analyses, ~120 genomic DNAs were screened by hybridization to a conserved region within the most recently formed HML-2 proviruses. Variable fragments were observed, as interpreted to represent individual HML-2 junction sites, that were detected in as few as 1 or 2 samples within the entire analyzed set. For example, the fragment in lane 25 of the CPSII control group at ~1400 bp was not observed for any other sample, and within all samples, the frequency is ~0.008, or 1 in 120 tested DNAs (asterisked, and indicated with arrowhead in Figure 4.3). Within the

SMRI unblots, other fragments with similarly low frequencies were observed (arrowheads in Figure 4.4 and 4.5). Such low frequencies would indicate very recent formation, and possibly conserved function.

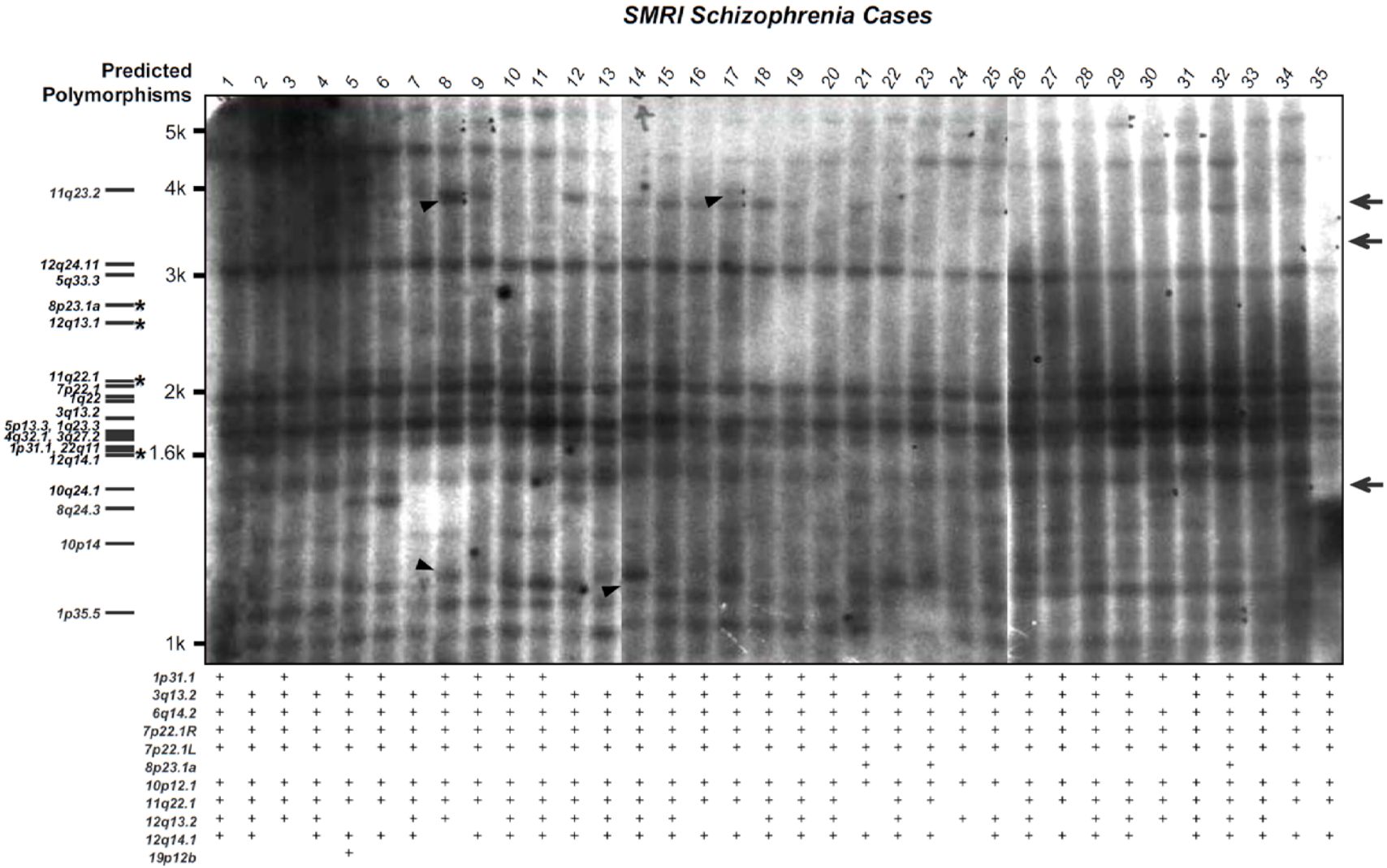


Figure 4.4. The case distribution of polymorphic HML-2 proviruses in schizophrenia. Samples were unblotted according to group with a HML-2-specific oligonucleotide to visualize polymorphic HML-2. Shown are case group hybridization. Black Arrowheads were used to indicate observed low frequency fragments. Labeling is as in Figure 4.3.

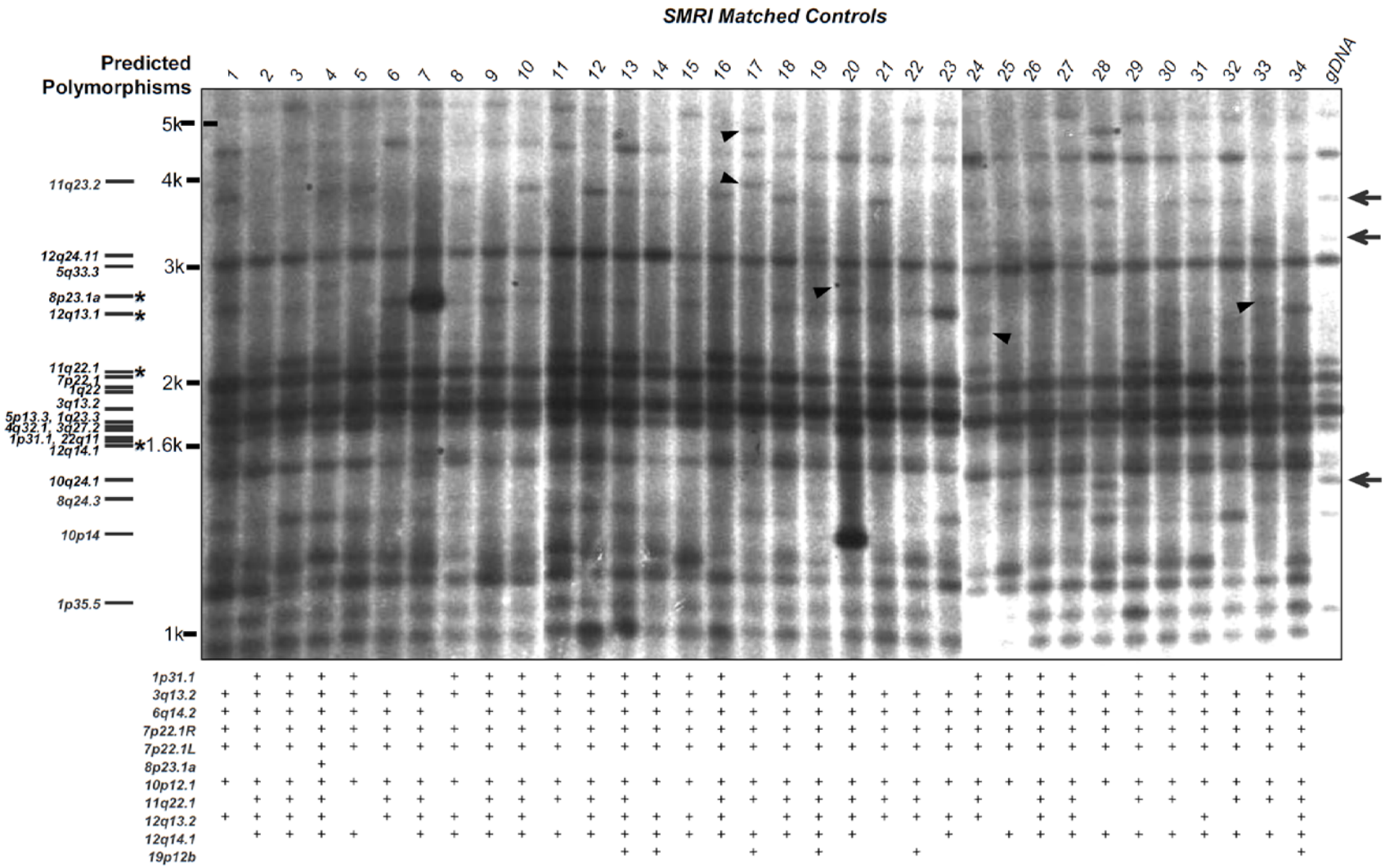


Figure 4.5. The control distribution of polymorphic HML-2 proviruses in schizophrenia. Samples were unblotted according to group with a HML-2-specific oligonucleotide to visualize polymorphic HML-2. Shown are hybridizations of the control group, matched for age, race, and sex. Arrowheads were used to indicate observed low frequency fragments. Labeling is as in Figure 4.3. The added gDNA sample was genomic DNA from the T47D cell line.

Discussion

In this study we analyzed the distribution and frequencies of polymorphic HML-2 proviruses within genomic DNAs representative of two human diseases. Fifty DNA samples were from DNAs from subsequently diagnosed breast cancer patients and individuals with no subsequent history [179], and 70 samples were from clinically diagnosed schizophrenics and matched undiagnosed controls [175]. For these purposes, we utilized a high-resolution DNA hybridization technique, unblotting, in order to infer the case-control distribution of putative 'novel' polymorphic HML-2, for which the integration sites have not yet been described. To our knowledge, this is the first report of such a comparison, and the largest representative set of genomic DNAs with inference of uncharacterized polymorphic proviruses from the HML-2 group.

Comparing the unblotted WGA-DNAs to the *in silico* predictions based on the published genome build allowed us to predict the analogous HML-2 loci for at least four sites (*in silico* bands are asterisked in Figure 4.3). PCR analysis provided support for four of these sites. The fragments predicted to represent the 11q22.1 element were the most clear, at ~2.1kb, and comparison to the PCR data further supported the verification of the provirus. Near the predicted size of ~1.6 kb were fragments that were consistent with the sample distribution of the 12q14.1 provirus. The 12q13.2 and K115 fragments are around 2.2kb and 2.4kb, respectively, and likewise, the *in silico* and PCR comparisons were also in agreement. Further support was offered in results from a separate study, in which we were able to amplify the 5'LTRs belonging to the 11q22.1 and 12q13.2 elements from elutes of their corresponding hybridized fragments (data not shown). Thus, we could conclusively assign those fragments to their respective HML-2 proviruses.

The expected patterns for the remaining polymorphic loci which have been described were not observed, although most of the elements were present in higher frequencies, were fixed within the sample set, or not predicted to be observed following *Bsr*I digestion. Although the 1p31.1 element could not be discerned, the junction fragment for this particular provirus is estimated around ~1.6 kb, and we speculate its 'masking' within other hybridized fragments of similar size, given the robust signals and *in silico* predictions corresponded to multiple HML-2 proviruses near each site. However, the apparent absence could be due to a shared mutation within a predicted *Bsr*I restriction site or introduced into the 5'LTR. Excluding the 1p31.1 element, the remaining detected HML-2 polymorphic elements were consistent in predicted fragment lengths and sample distribution, providing support for our strategy.

Across all samples, the patterns of HML-2 proviruses we observed in unblotting were mostly in agreement with the *in silico* restriction analysis of the published genome sequence. Also, the PCR analysis of 11 described polymorphic elements verified the fragments of four polymorphic integrations, each of which was also consistent with *in silico* predictions. From the total number of fragments that were inconsistent in pattern with *in silico* or PCR analyses, we were able to tentatively compare the frequencies of several bands between groups, excluding regions in each unblot for which the detected HML-2 fragments were difficult to resolve (indicated in Figure 4.3). The case-control frequencies for 5 polymorphic elements were compared between the CPSII sample groups, and 2 between either SMRI group, none of which were found to have significant difference. Collectively, these data suggest that for the

detectable ‘new’ polymorphic integrations within this sample set, no single element appears to be associated with an increased risk for either disease.

The K115 and K113 proviruses were the first insertionally polymorphic HML-2 to be discovered [95]. In multiple reports, specific attention has been given to both proviruses as possible candidates for roles in human diseases, based on the properties of relatively recent germline integration (estimated at <200,000 years and ~1.2 mya, respectively) and presence in a fraction of the population [95]. In analyses of the K113 and K115 elements alone, attempts have been made to determine if there is a genetic association of either provirus with one of several human diseases, including breast cancer [205], multiple sclerosis [209], schizophrenia [167], seminomas [210], and a few autoimmune diseases [96, 204]. In most such cases, association of either provirus with disease has been rejected; although one study found an increase in the frequency of K113 in multiple sclerosis [96], the results were not replicated in a larger study [209]. Similarly, our initial observations of a higher prevalence of K115 to breast cancer cases was of interest, were not replicated. Given the outcome of the PCR analysis of the larger sample set, the necessity for such added analyses is made clear. It would be speculative to draw conclusions from the observed differences of 11q22.1 or K113 within the SMRI DNAs, however given previous studies [167], and the relatively high frequency of the 11q22.1 provirus, replication of such a difference may be unlikely.

In studies which K113 and K115 have been analyzed for associations with human disease, their detectable frequencies have ranged from ~10-20% for K113 and ~5-12% for K115 (for example see [96, 210]). Our results are consistent with these observations, with the exception of the K115 provirus in ~24% of cases in the initial

screen (Table 4.2). A frequency of 24% of tested individuals for this particular provirus is not completely uncommon, as the detection has been even as high as >40% depending on ethnicity of the samples tested [95, 96]. Similarly, K113 has been observed to levels of ~30%, again with reference to ethnicity [95, 96]. Given the variance with respect to race, the observed frequencies of the K115 provirus among DNAs from breast cancer cases may be a consequence of an uneven racial representation of the sample set. Alternatively, the higher frequency of K115 we observed in cases could be due to stochastic effects from the relatively small sample size used for the present analysis. As the samples were de-identified, we were only provided the case or control identities per sample. Thus, we can only speculate on the factors in the observed distribution.

To date, all reports that have attempted to detect a genetic association of individual HML-2 and a disease risk have offered little support for any implications in disease. A provirus that did have negative effects to the host would likely be removed from the population under a purifying selection, or may be detectable at low levels within the host species. Detection of such an element may necessitate larger sample sizes than have been used in approaches searching for a genetic association and would require the detection of the provirus. Also, conclusive proof of a genetic association would be required in the detection of clonal integrations, however as has been stressed several times, no novel proviruses have yet been identified, nor has any single polymorphic provirus been associated with any disease. Repeated searches for a disease association with one or two particular elements alone, such as has been the case for K113 and K115, will likely have similar outcomes as have been observed. We attempted to overcome such limitations by screening genomic DNAs in a highly specific DNA hybridization from two

different diseases in a case-control comparison, and though we interpret the data to indicate the detectable presence of several as-yet-uncharacterized polymorphic proviruses, there is no offered statistical support for a genetic association to either disease.

There are alternative explanations for the observed banding patterns. For example, the observed fragments could possibly have been a consequence of base changes within the target sequence for the *Bsr*I restriction enzyme. Such mutations could result in unexpected fragments of the representative HML-2 proviruses in some samples, giving the impression of a polymorphic integration. Also, mutation could lead to the generation of a new restriction site, for example within the 5' LTR, that would prevent the detection of the corresponding junction fragment by the probe. We searched for such an example from the fragments that were tentatively identified as described HML-2 (asterisked in Figures 4.3 to 4.5), and found the PCR and unblot data were in agreement, suggesting the *Bsr*I target sites for these particular elements have not been disrupted. However, it is difficult to exclude the possibility of mutation having occurred at the restriction sites for other detected proviruses (or possible shared SNPs among subjects) without sequence information for each band.

Nevertheless, we have provided evidence of a number of polymorphic proviruses that vary in frequencies among the samples tested here, some of which are present at quite low frequencies (arrowheads in Figures 4.3 to 4.5, and for specific examples, also refer to lanes 17 and 28 of Figure 4.5 and lane 25 of the control group in Figure 4.3). For the ~120 genomic DNAs, between 18 and 22 bands were observed per sample About 10-15 fragments were observed for which the corresponding described

provirus could not be inferred from *in silico* or PCR analyses. Given the sample size, it is likely that some of the observed bands represent recently integrated proviruses which are present in just a portion of individuals. For those fragments detected in few individuals, the estimated frequencies are below ~1% of the total number of samples, a far lower representation than seen in any other polymorphic provirus. Such a provirus would be highly conserved, and we speculate might also exhibit retained functions. Future work on the project will likely shed light on these issues.

Chapter 5

Conclusions and Future Directions

Conclusions

Human endogenous retroviruses (HERVs) result from the integration of retroviral DNA into germline cells. Our analysis of the HERV-K(HML-2) group has revealed their representation in the published genome by ~90 full-length proviruses. The HML-2 proviruses can be divided into three subgroups, based on the phylogenetic comparison of the paired LTRs belonging to each element, namely, the LTR5Hs, LTR5A, and LTR5B. The analysis of derived subgroup consensus sequences has allowed the inference of each subgroup by shared, specific sequence motifs present within each subgroup, and the identification of each element by such subgroup corresponds to their phylogeny. The LTR5B subgroup represents the oldest elements, including a few shared sites in the Old World Monkeys, and none of which have been specific to humans. The phylogenetic analysis of each subgroup also reveals that the LTR5B proviruses are ancestral to both the LTR5A and the LTR5-Hs elements, which is supported through the estimated times of activities for proviruses belonging to either subgroup. Of the two, only members of the LTR5-Hs have been active in direct lineage leading to humans.

Our results lend clarification of the HML-2 present within the available genome sequences, and provide more detail to the understanding of their genome context and evolutionary history. As inferred from the phylogenetic analysis of each subgroup, the evolutionary history of the LTR5-Hs has differed markedly from that of the LTR5A and LTR5B subgroups. The latter two groups have experienced bursts of increase in copy number as the result of association with segmental duplication events within the host. Also we have presented findings that the LTR5A elements are associated with a specific group of genome segments, and are found in more than half of the newly formed

duplicons. Furthermore, additional segments were identified for which members of the LTR5-Hs were observed to have integrated, in one case just a few kb from the corresponding integration site of the LTR5A duplicates, and in the other case in the proximal integration of the K115 provirus. The significance of these observations is not clear, however we speculate there is a certain propensity of integrations from the HML-2 group within such duplications. Given the identification of the K115 element at such a site, the possibility exists that the K115 provirus itself may experience a similar fate as the LTR5A duplicates, given time. For either LTR5A or B subgroups, the duplicated elements account for nearly one third of the respective subgroup, indicating even less frequent germline infection of either subgroup. In contrast are the recent HML-2, including human-specific and polymorphic integrations, which have originated predominantly from unique integrations of exogenous viruses.

The HML-2 group represents the only HERVs with human-specific integrations, of which at least 11 HML-2 proviruses are polymorphic in integration site and frequency, and 5 have not yet been fixed within the population. While the vast majority of HERVs have been bombarded with mutations, including insertions, deletions, truncations, and also solo-LTR formations, several members of the HML-2 group are observed with intact open reading frames in at least some or all genes, and with the capacity to express functional proteins. However, an infectious provirus from the HML-2 group has not been observed. All HML-2-encoded genes, including the *env* reading frame, are highly conserved and exhibit signs of purifying selection, an observation that indicates entry into the germline by the exogenous infection. Also, the rate of germline formation appears to have been constant following the divergence from chimpanzee.

Taken together these observations suggest that replication competent HML-2 proviruses exist at unfixed sites within a subset of the human population, the most likely candidates for which are rare, recently integrated proviruses.

Members of the LTR5-Hs, and particularly those with species-specificity to humans, and especially those that are polymorphic within humans, have been implicated in a number of human diseases. Most predominant have been the reports of expression in several types of cancer, and within neurodegenerative and autoimmune disorders. Despite the mounting evidence suggesting the expression from HML-2 may have a clinically significant role in disease, the reason(s) for the aberrant expression of HML-2 remain unclear. Moreover, all attempts to infer the genetic association of a polymorphic provirus have been rejected, and no novel (or clonal) provirus has been detected in this respect. Our analyses also offer no statistical support for a genetic association with disease. Furthermore, as inferred from the HML-2-specific hybridizations by unblot, our analysis also offers no support for a 'novel' polymeric provirus in association with disease. Nevertheless, we provide evidence for at least ~10-15 uncharacterized polymorphic proviruses belonging to the HML-2 group. We observed such bands in varying frequencies, with a few bands present in more than half of the samples tested, but also were a few hybridized fragments present in just one or two samples (less than 1%), and we tentatively interpret these data to indicate restriction fragments containing the junction sites of undescribed proviruses present in low frequencies within humans.

We have previously attempted several PCR approaches in order to sequence individual fragments either as directly excised from unblotted agarose, or from total bulk WGA-DNAs. These techniques included, but were not limited to a linear-amplification-mediated (LAM) PCR, a linker-mediated (LM) PCR, and inverse PCR (data not shown). To identify the chromosomal locations of such proviruses from selected genomic DNAs, we recently designed and attempted a combined strategy of PCR amplification and deep sequencing using WGA-DNA as a template. Our initial analyses have indicated about 10-20 possible sites of interest, with reference to the most recent genome build, and we have begun preliminary PCR screening of the WGA-DNAs included in the original Illumina library (data not shown). We anticipate these data will generate a few confirmed sites, and may lead to the identification of novel polymorphic proviruses with possible function in future work.

REFERENCES

References

1. McClintock B: **Intranuclear systems controlling gene action and mutation.** *Brookhaven Symp Biol* 1956**:**58-74.
2. IHGSC: **A physical map of the human genome.** *Nature* 2001, **409:**934-941.
3. Cordaux R, Batzer MA: **The impact of retrotransposons on human genome evolution.** *Nat Rev Genet* 2009, **10:**691-703.
4. Hartl DL, Lohe AR, Lozovskaya ER: **Regulation of the transposable element mariner.** *Genetica* 1997, **100:**177-184.
5. Craig NL, Craigie R, Gellert M, Lambowitz AM: **Transposition Reactions that Involve Only DNA Intermediates.** In *Mobile DNA II*. Edited by Craig NL. Eashington, D.C.: ASM Press; 2002: 304-565
6. Pace JK, 2nd, Feschotte C: **The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage.** *Genome Res* 2007, **17:**422-432.
7. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110:**462-467.
8. Kaplan N, Darden T, Langley CH: **Evolution and extinction of transposable elements in Mendelian populations.** *Genetics* 1985, **109:**459-480.
9. Silva JC, Kidwell MG: **Evolution of P elements in natural populations of Drosophila willistoni and D. sturtevanti.** *Genetics* 2004, **168:**1323-1335.

10. Houck MA, Clark JB, Peterson KR, Kidwell MG: **Possible horizontal transfer of Drosophila genes by the mite Proctolaelaps regalis.** *Science* 1991, **253:**1125-1128.

11. Neil S, Bieniasz P: **Human immunodeficiency virus, restriction factors, and interferon.** *J Interferon Cytokine Res* 2009, **29:**569-580.

12. Boeke JD, Stoye JP: **Retrotransposons, endogenous retroviruses, and the evolution of retroelements.** In *Retroviruses.* Edited by Coffin J, Hughes S, Varmus H. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997: 343-435

13. Baltimore D: **RNA-dependent DNA polymerase in virions of RNA tumour viruses.** *Nature* 1970, **226:**1209-1211.

14. Temin HM, Mizutani S: **RNA-dependent DNA polymerase in virions of Rous sarcoma virus.** *Nature* 1970, **226:**1211-1213.

15. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution.** *Nature* 1980, **284:**601-603.

16. Orgel LE, Crick FH: **Selfish DNA: the ultimate parasite.** *Nature* 1980, **284:**604-607.

17. Kazazian HH, Jr.: **Mobile elements: drivers of genome evolution.** *Science* 2004, **303:**1626-1632.

18. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH, Jr.: **High frequency retrotransposition in cultured mammalian cells.** *Cell* 1996, **87:**917-927.

19. Kazazian HH, Jr.: **An estimated frequency of endogenous insertional mutations in humans.** *Nat Genet* 1999, **22:**130.

20. Salem AH, Kilroy GE, Watkins WS, Jorde LB, Batzer MA: **Recently integrated Alu elements and human genomic diversity.** *Mol Biol Evol* 2003, **20:**1349-1361.

21. Vincent BJ, Myers JS, Ho HJ, Kilroy GE, Walker JA, Watkins WS, Jorde LB, Batzer MA: **Following the LINEs: an analysis of primate genomic variation at human-specific LINE-1 insertion sites.** *Mol Biol Evol* 2003, **20:**1338-1348.

22. Katzourakis A, Tristem M: **Phylogeny of Human Endogenous and Exogenous Retroviruses.** In *Retroviruses and Primate Genome Evolution.* Edited by Sverdlov EV. Georgetown, TX: Landes Biosciences; 2005: 186-203

23. Glazko GV, Nei M: **Estimation of divergence times for major lineages of primate species.** *Mol Biol Evol* 2003, **20:**424-434.

24. Karlsson H, Schroder J, Bachmann S, Bottmer C, Yolken RH: **HERV-W-related RNA detected in plasma from individuals with recent-onset schizophrenia or schizoaffective disorder.** *Mol Psychiatry* 2004, **9:**12-13.

25. Frank O, Giehl M, Zheng C, Hehlmann R, Leib-Mosch C, Seifarth W: **Human endogenous retrovirus expression profiles in samples from brains of patients with schizophrenia and bipolar disorders.** *J Virol* 2005, **79:**10890-10901.

26. Nakamura A, Okazaki Y, Sugimoto J, Oda T, Jinno Y: **Human endogenous retroviruses with transcriptional potential in the brain.** *J Hum Genet* 2003, **48:**575-581.

27. Karlsson H, Bachmann S, Schroder J, McArthur J, Torrey EF, Yolken RH: **Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia.** *Proc Natl Acad Sci U S A* 2001, **98:**4634-4639.

28. Voisset C, Andrawiss M: **Retroviruses at a glance.** *Genome Biology* 2000, **1**.

29. Vogt P: **Historical introduction to the general properties of retroviruses.** In *Retroviruses.* Edited by Coffin JM, Hughes SH, Varmus H. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997: 1-25

30. Coffin JM: **Structure and classification of retroviruses.** In *The Retroviridae.* Edited by Levy J. New York: Plenum Press; 1994: 19-49

31. Vogt V: **Retroviral virions and genomes.** In *Retroviruses.* Edited by Coffin J, Hughes S, Varmus H. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997

32. Jern P, Sperber GO, Blomberg J: **Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy.** *Retrovirology* 2005, **2:**50.

33. Blomberg J, Benachenhou F, Blikstad V, Sperber G, Mayer J: **Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations.** *Gene* 2009, **448:**115-123.

34. Katzourakis A, Tristem M, Pybus OG, Gifford RJ: **Discovery and analysis of the first endogenous lentivirus.** *Proc Natl Acad Sci U S A* 2007, **104:**6261-6265.

35. Gilbert C, Maxfield DG, Goodman SM, Feschotte C: **Parallel germline infiltration of a lentivirus in two Malagasy lemurs.** *PLoS Genet* 2009, **5:**e1000425.

36. Gifford RJ, Katzourakis A, Tristem M, Pybus OG, Winters M, Shafer RW: **A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution.** *Proc Natl Acad Sci U S A* 2008, **105:**20362-20367.

37. Gifford R, Tristem M: **The evolution, distribution and diversity of endogenous retroviruses.** *Virus Genes* 2003, **26:**291-315.

38. Bannert N, Kurth R: **The evolutionary dynamics of human endogenous retroviral families.** *Annu Rev Genomics Hum Genet* 2006, **7:**149-173.

39. Mayer J, Blomberg J, Seal RL: **A revised nomenclature for transcribed human endogenous retroviral loci.** *Mob DNA*, **2:**7.

40. Rabson AB, Graves BJ: **Synthesis and processing of viral RNA.** In *Retroviruses.* Edited by Coffin J, Hughes S, Varmus H. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997

41. Telesnitsky A, Goff SP: **Reverse transcriptase and the generation of retroviral DNA.** In *Retroviruses.* Edited by Coffin J, Hughes S, Varmus H. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997

42. Freed EO: **HIV-1 replication.** *Somat Cell Mol Genet* 2001, **26:**13-33.

43. Rein A, Datta SA, Jones CP, Musier-Forsyth K: **Diverse interactions of retroviral Gag proteins with RNAs.** *Trends Biochem Sci*.

44. Swanstrom R, WIlls JW: **Synthesis, assembly, and processing of viral proteins.** In *Retroviruses.* Edited by Coffin J, Hughes S, Varmus H. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997

45. Nowotny M: **Retroviral integrase superfamily: the structural perspective.** *EMBO Rep* 2009, **10:**144-151.

46. Hunter E: **Viral entry and receptors.** In *Retroviruses.* Edited by Coffin J, Hughes S, Varmus H. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997

47. Nisole S, Saib A: **Early steps of retrovirus replicative cycle.** *Retrovirology* 2004, **1:**9.

48. Temin HM: **Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation.** *Proc Natl Acad Sci U S A* 1993, **90:**6900-6903.

49. Coffin JM: **Genetic diversity and evolution of retroviruses.** *Curr Top Microbiol Immunol* 1992, **176:**143-164.

50. Suzuki Y, Craigie R: **The road to chromatin - nuclear entry of retroviruses.** *Nat Rev Microbiol* 2007, **5**.

51. Brown P: **Integration.** In *Retroviruses.* Edited by Coffin J, Hughes S, Varmus H. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997

52. Katz RA, Greger JG, Boimel P, Skalka AM: **Human immunodeficiency virus type 1 DNA nuclear import and integration are mitosis independent in cycling cells.** *J Virol* 2003, **77:**13412-13417.

53. Hare S, Gupta SS, Valkov E, Engelman A, Cherepanov P: **Retroviral intasome assembly and inhibition of DNA strand transfer.** *Nature*, **464:**232-236.

54. Maertens GN, Hare S, Cherepanov P: **The mechanism of retroviral integration from X-ray structures of its key intermediates.** *Nature*, **468:**326-329.

55. Craigie R: **When four became one.** *Nature* 2010, **464**.

56. Desfarges S, Ciuffi A: **Retroviral integration site selection.** *Viruses* 2010, **2**.

57. Mager DL: **Polyadenylation function and sequence variability of the long terminal repeats of the human endogenous retrovirus-like family RTVL-H.** *Virology* 1989, **173:**591-599.

58. Toszer J: **Comparative studies on retroviral proteases: substrate specificity.** *Viruses* 2010, **2:**147-165.

59. Dorfman T, Luban J, Goff SP, Haseltine WA, Gottlinger HG: **Mapping of functionally important residues of a cysteine-histidine box in the human immunodeficiency virus type 1 nucleocapsid protein.** *J Virol* 1993, **67:**6159-6169.

60. Zhang Y, Barklis E: **Nucleocapsid protein effects on the specificity of retrovirus RNA encapsidation.** *J Virol* 1995, **69:**5716-5722.

61. Herniou E, Martin J, Miller K, Cook J, Wilkinson M, Tristem M: **Retroviral diversity and distribution in vertebrates.** *J Virol* 1998, **72:**5955-5966.

62. Stoye J, Coffin J: **The Molecular Biology of Retroviruses.** In *Endogenous Retroviruses.* 2nd edition. Edited by RA W, N T, HE V, JM C. New York, NY: CSHL Press; 1985: 357-404

63. Johnson WE, Coffin JM: **Constructing primate phylogenies from ancient retrovirus sequences.** *Proc Natl Acad Sci* 1999, **96:**10254-10260.

64. Hughes JF, Coffin JM: **Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution.** *Nat Genet* 2001, **29:**487-489.

65. Hughes JF, Coffin JM: **Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome.** *Genetics* 2005, **171:**1183-1194.

66. Boeke J, Stoye J: **Retrotransposons, endogenous retroviruses, and the evolution of retroelements.** In *Retroviruses.* Edited by Coffin J, Hughes S, Varmus H. New York, NY: CSHL Press; 1997: 343-435

67. Medstrand P, Mager DL: **Human-specific integrations of the HERV-K endogenous retrovirus family.** *J Virol* 1998, **72:**9782-9787.

68. Tristem M: **Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database.** *J Virol* 2000, **74:**3715-3730.

69. Enard W, Paabo S: **Comparative primate genomics.** *Annu Rev Genomics Hum Genet* 2004, **5:**351-378.

70. Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M: **High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection.** *Mol Biol Evol* 2005, **22:**814-817.

71. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M: **Long-term reinfection of the human genome by endogenous retroviruses.** *Proc Natl Acad Sci* 2004, **101:**4894-4899.

72. Jern P, Sperber GO, Blomberg J: **Definition and variation of human endogenous retrovirus H.** *Virology* 2004, **327:**93-110.

73. Sverdlov EV: **Retroviruses and primate evolution.** *BioEssays* 2000, **22**.

74. Mayer J, Meese EU: **The human endogenous retrovirus family HERV-K(HML-3).** *Genomics* 2002, **80**.

75. Jern P, Coffin JM: **Effects of retroviruses on host genome function.** *Annu Rev Genet* 2008, **42:**709-732.

76. Stoye JP: **Endogenous retroviruses: Still active after all these years?** *Current Biology* 2001, **11:**R914-R916.

77. Cohen CJ, Lock WM, Mager DL: **Endogenous retroviral LTRs as promoters for human genes: a critical assessment.** *Gene* 2009, **448:**105-114.

78. Medstrand P, van de Lagemaat LN, Mager DL: **Retroelement distributions in the human genome: variations associated with age and proximity to genes.** *Genome Res* 2002, **12:**1483-1495.

79. Smit AFA: **Interspersed repeats and other mementos of transposable elements in mammalian genomes.** *Current Opinion in Genetics & Development* 1999, **9:**657-663.

80. Gogvadze E, Stukacheva E, Buzdin A, Sverdlov E: **Human specific modulation of transcriptional activity provided by endogenous retroviral inserts.** *J Virol* 2009.

81. Mager DL, van de Lagemaat LN, Medstrand P: **Genomic Distributions of Human Retroelements.** In *Retrovriuses and Primate Genome Evolution.* Edited by Sverdlov E. Gergetown, TX: Landes Biosciences; 2005: 104-122

82. van de Lagemaat LN, Medstrand P, Mager DL: **Multiple effects govern endogenous retrovirus survival patterns in human gene introns.** *Genome Biol* 2006, **7:**R86.

83. Vinogradova T, Volik S, Lebedev Y, Shevchenko Y, Lavrentyeva I, Khil P, Grzeschik KH, Ashworth LK, Sverdlov E: **Positioning of 72 potentially full size LTRs of human endogenous retroviruses HERV-K on the human chromosome 19 map. Occurrences of the LTRs in human gene sites.** *Gene* 1997, **199:**255-264.

84. Kurdyukov SG, Lebedev YB, Artamonova, II, Gorodentseva TN, Batrak AV, Mamedov IZ, Azhikina TL, Legchilina SP, Efimenko IG, Gardiner K, Sverdlov ED: **Full-sized HERV-K (HML-2) human endogenous retroviral LTR sequences on human chromosome 21: map locations and evolutionary history.** *Gene* 2001, **273:**51-61.

85. Buzdin A, Kovalskaya-Alexandrova E, Gogvadze E, Sverdlov E: **At least 50% of human-specific HERV-K (HML-2) long terminal repeats serve in vivo as active promoters for host nonrepetitive DNA transcription.** *J Virol* 2006, **80:**10752-10762.

86. Konkel MK, Batzer MA: **A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome.** *Semin Cancer Biol*, **20:**211-221.

87. Katzourakis A, Pereira V, Tristem M: **Effects of recombination rate on human endogenous retrovirus fixation and persistence.** *J Virol* 2007, **81:**10712-10717.

88. Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, Burt A, Tristem M: **Rate of recombinational deletion among human endogenous retroviruses.** *J Virol* 2007, **81:**9437-9442.

89. Hughes JF, Coffin JM: **Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution.** *Proc Natl Acad Sci* 2004, **101:**1668-1672.

90. Reus K, Mayer J, Sauter M, Scherer D, Muller-Lantzsch N, Meese E: **Genomic organization of the human endogenous retrovirus HERV-K(HML-2.HOM) (ERVK6) on chromosome 7.** *Genomics* 2001, **72:**314-320.

91. Jern P, Coffin JM: **Host-retrovirus arms race: trimming the budget.** *Cell Host Microbe* 2008, **4:**196-197.

92. Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP: **Gene conversion: mechanisms, evolution and human disease.** *Nat Rev Genet* 2007, **8:**762-775.

93. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M: **Long-term reinfection of the human genome by endogenous retroviruses.** *Proc Natl Acad Sci U S A* 2004, **101:**4894-4899.

94. Macfarlane C, Simmonds P: **Allelic variation of HERV-K(HML-2) endogenous retroviral elements in human populations.** *J Mol Evol* 2004, **59:**642-656.

95. Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J: **Insertional polymorphisms of full-length endogenous retroviruses in humans.** *Curr Biol* 2001, **11:**1531-1535.

96. Moyes DL, Martin A, Sawcer S, Temperton N, Worthington J, Griffiths DJ, Venables PJ: **The distribution of the endogenous retroviruses HERV-K113 and HERV-K115 in health and disease.** *Genomics* 2005, **86:**337-341.

97. Hedges DJ, Deininger PL: **Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity.** *Mutat Res* 2007, **616:**46-59.

98. Kamp C, Hirschmann P, Voss H, Huellen K, Vogt PH: **Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination events.** *Hum Mol Genet* 2000, **9:**2563-2572.

99. Leib-Mosch C, Seifarth W, Schon U: **Influence of Human Endogenous Retroviruses on Cellular Gene Expression.** In *Retroviruses and Primate Genome Evolution.* Edited by Sverdlov EV. Georgetown, TX: Landes Bioscience; 2005: 123-143

100. Samuelson LC, Wiebauer K, Snow CM, Meisler MH: **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:**2513-2520.

101. Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH: **Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene.** *Genes Dev* 1992, **6:**1457-1465.

102. Medstrand P, Landry JR, Mager DL: **Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans.** *J Biol Chem* 2001, **276:**1896-1903.

103. Bi S, Gavrilova O, Gong DW, Mason MM, Reitman M: **Identification of a placental enhancer for the human leptin gene.** *J Biol Chem* 1997, **272:**30583-30588.

104. Kapitonov VV, Jurka J: **The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor.** *J Mol Evol* 1999, **48:**248-251.

105. Schulte AM, Lai S, Kurtz A, Czubayko F, Riegel AT, Wellstein A: **Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus.** *Proc Natl Acad Sci U S A* 1996, **93:**14759-14764.

106. Czubayko F, Schulte AM, Berchem GJ, Wellstein A: **Melanoma angiogenesis and metastasis modulated by ribozyme targeting of the secreted growth factor pleiotrophin.** *Proc Natl Acad Sci U S A* 1996, **93:**14753-14758.

107. Schulte AM, Malerczyk C, Cabal-Manzano R, Gajarsa JJ, List HJ, Riegel AT, Wellstein A: **Influence of the human endogenous retrovirus-like element HERV-E.PTN on the expression of growth factor pleiotrophin: a critical role of a retroviral Sp1-binding site.** *Oncogene* 2000, **19:**3988-3998.

108. Blond JL, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes S, Mandrand B, Mallet F, Cosset FL: **An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor.** *J Virol* 2000, **74:**3321-3329.

109. Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, Lucotte G, Duret L, Mandrand B: **The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology.** *Proc Natl Acad Sci* 2004, **101:**1731-1736.

110. Heidmann O, Vernochet C, Dupressoir A, Heidmann T: **Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals.** *Retrovirology* 2009, **6:**107.

111. Stoye JP: **Proviral protein provides placental function.** *Proc Natl Acad Sci U S A* 2009, **106:**11827-11828.

112. Bjerregaard B, Holck S, Christensen IJ, Larsson LI: **Syncytin is involved in breast cancer-endothelial cell fusions.** *Cell Mol Life Sci* 2006, **63:**1906-1911.

113. Wray GA: **The evolutionary significance of cis-regulatory mutations.** *Nat Rev Genet* 2007, **8:**206-216.

114. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20:**1377-1419.

115. Buzdin A, Ustyugova S, Khodosevich K, Mamedov I, Lebedev Y, Hunsmann G, Sverdlov E: **Human-specific subfamilies of HERV-K (HML-2) long terminal repeats: three master genes were active simultaneously during branching of hominoid lineages.** *Genomics* 2003, **81:**149-156.

116. Kovalskaya E, Buzdin A, Gogvadze E, Vinogradova T, Sverdlov E: **Functional human endogenous retroviral LTR transcription start sites are located between the R and U5 regions.** *Virology* 2006, **346:**373-378.

117. Callahan R, Drohan W, Tronick S, Schlom J: **Detection and cloning of human DNA sequences related to the mouse mammary tumor virus genome.** *Proc Natl Acad Sci U S A* 1982, **79:**5503-5507.

118. Ono M: **Molecular cloning and long terminal repeat sequences of human endogenous retrovirus genes related to types A and B retrovirus genes.** *J Virol* 1986, **58:**937-944.

119. Ono M, Yasunaga T, Miyata T, Ushikubo H: **Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome.** *J Virol* 1986, **60:**589-598.

120. Romano CM, Ramalho RF, Zanotto PM: **Tempo and mode of ERV-K evolution in human and chimpanzee genomes.** *Arch Virol* 2006, **151:**2215-2228.

121. Magin C, Lower R, Lower J: **cORF and RcRE, the Rev/Rex and RRE/RxRE homologues of the human endogenous retrovirus family HTDV/HERV-K.** *J Virol* 1999, **73:**9496-9507.

122. Armbruester V, Sauter M, Krautkraemer E, Meese E, Kleiman A, Best B, Roemer K, Mueller-Lantzsch N: **A novel gene from the human endogenous retrovirus K expressed in transformed cells.** *Clin Cancer Res* 2002, **8:**1800-1807.

123. Mayer J, Ehlhardt S, Seifert M, Sauter M, Muller-Lantzsch N, Mehraein Y, Zang KD, Meese E: **Human endogenous retrovirus HERV-K(HML-2) proviruses**

with Rec protein coding capacity and transcriptional activity. *Virology* 2004, **322:**190-198.

124. Magin-Lachmann C, Hahn S, Strobel H, Held U, Lower J, Lower R: **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:**10359-10371.

125. Armbruester V, Sauter M, Roemer K, Best B, Hahn S, Nty A, Schmid A, Philipp S, Mueller A, Mueller-Lantzsch N: **Np9 protein of human endogenous retrovirus K interacts with ligand of numb protein X.** *J Virol* 2004, **78:**10310-10319.

126. Denne M, Sauter M, Armbruester V, Licht JD, Roemer K, Mueller-Lantzsch N: **Physical and functional interactions of human endogenous retrovirus proteins Np9 and rec with the promyelocytic leukemia zinc finger protein.** *J Virol* 2007, **81:**5607-5616.

127. Mayer J, Meese EU: **Presence of dUTPase in the various human endogenous retrovirus K (HERV-K) families.** *J Mol Evol* 2003, **57:**642-649.

128. Knossl M, Lower R, Lower J: **Expression of the human endogenous retrovirus HTDV/HERV-K is enhanced by cellular transcription factor YY1.** *J Virol* 1999, **73:**1254-1261.

129. Ono M, Kawakami M, Ushikubo H: **Stimulation of expression of the human endogenous retrovirus genome by female steroid hormones in human breast cancer cell line T47D.** *J Virol* 1987, **61:**2059-2062.

130. Fuchs NV, Kraft M, Tondera C, Hanschmann KM, Lower J, Lower R: **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*, **85:**3436-3448.

131. Belshaw R, Dawson AL, Woolven-Allen J, Redding J, Burt A, Tristem M: **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:**12507-12514.

132. Dewannieux M, Blaise S, Heidmann T: **Identification of a functional envelope protein from the HERV-K family of human endogenous retroviruses.** *J Virol* 2005, **79:**15573-15577.

133. Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, Pierron G, Heidmann T: **Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements.** *Genome Res* 2006, **16:**1548-1556.

134. Ruprecht K, Ferreira H, Flockerzi A, Wahl S, Sauter M, Mayer J, Mueller-Lantzsch N: **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:**10008-10016.

135. Beimforde N, Hanke K, Ammar I, Kurth R, Bannert N: **Molecular cloning and functional characterization of the human endogenous retrovirus K113.** *Virology* 2008, **371:**216-225.

136. Boller K, Schonfeld K, Lischer S, Fischer N, Hoffmann A, Kurth R, Tonjes RR: **Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles.** *J Gen Virol* 2008, **89:**567-572.

137. Lee YN, Bieniasz PD: **Reconstitution of an infectious human endogenous retrovirus.** *PLoS Pathog* 2007, **3:**e10.

138. Rosenberg N, Jolicoeur P: **Retroviral pathogenesis.** In *Retroviruses.* Edited by Coffin J, Hughes S, Varmus H. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997: 475-585

139. Li M, Huang X, Zhu Z, Gorelik E: **Sequence and insertion sites of murine melanoma-associated retrovirus.** *J Virol* 1999, **73:**9178-9186.

140. Mangeney M, Pothlichet J, Renard M, Ducos B, Heidmann T: **Endogenous retrovirus expression is required for murine melanoma tumor growth in vivo.** *Cancer Res* 2005, **65:**2588-2591.

141. Palmarini M, Sharp JM, de las Heras M, Fan H: **Jaagsiekte sheep retrovirus is necessary and sufficient to induce a contagious lung cancer in sheep.** *J Virol* 1999, **73:**6964-6972.

142. Caporale M, Cousens C, Centorame P, Pinoni C, De las Heras M, Palmarini M: **Expression of the jaagsiekte sheep retrovirus envelope glycoprotein is sufficient to induce lung tumors in sheep.** *J Virol* 2006, **80:**8030-8037.

143. Romanish MT, Cohen CJ, Mager DL: **Potential mechanisms of endogenous retroviral-mediated genomic instability in human cancer.** *Semin Cancer Biol*, **20:**246-253.

144. Frank O, Verbeke C, Schwarz N, Mayer J, Fabarius A, Hehlmann R, Leib-Mosch C, Seifarth W: **Variable transcriptional activity of endogenous retroviruses in human breast cancer.** *J Virol* 2008, **82:**1808-1818.

145. Seifarth W, Skladny H, Krieg-Schneider F, Reichert A, Hehlmann R, Leib-Mosch C: **Retrovirus-like particles released from the human breast cancer cell line T47-D display type B- and C-related endogenous retroviral sequences.** *J Virol* 1995, **69:**6408-6416.

146. Wang-Johanning F, Frost AR, Jian B, Epp L, Lu DW, Johanning GL: **Quantitation of HERV-K *env* gene expression and splicing in human breast cancer.** *Oncogene* 2003, **22:**1528-1535.

147. Wang-Johanning F, Frost AR, Johanning GL, Khazaeli MB, LoBuglio AF, Shaw DR, Strong TV: **Expression of human endogenous retrovirus k envelope transcripts in human breast cancer.** *Clin Cancer Res* 2001, **7:**1553-1560.

148. Contreras-Galindo R, Kaplan MH, Leissner P, Verjat T, Ferlenghi I, Bagnoli F, Giusti F, Dosik MH, Hayes DF, Gitlin SD, Markovitz DM: **Human endogenous retrovirus K (HML-2) elements in the plasma of people with lymphoma and breast cancer.** *J Virol* 2008, **82:**9329-9336.

149. Herbst H, Sauter M, Kuhler-Obbarius C, Loning T, Mueller-Lantzsch N: **Human endogenous retrovirus (HERV)-K transcripts in germ cell and trophoblastic tumours.** *Apmis* 1998, **106:**216-220.

150. Sauter M, Roemer K, Best B, Afting M, Schommer S, Seitz G, Hartmann M, Mueller-Lantzsch N: **Specificity of antibodies directed against Env protein of**

human endogenous retroviruses in patients with germ cell tumors. *Cancer Res* 1996, **56:**4362-4365.

151. Flockerzi A, Ruggieri A, Frank O, Sauter M, Maldener E, Kopper B, Wullich B, Seifarth W, Muller-Lantzsch N, Leib-Mosch C, et al: **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:**354.

152. Buscher K, Hahn S, Hofmann M, Trefzer U, Ozel M, Sterry W, Lower J, Lower R, Kurth R, Denner J: **Expression of the human endogenous retrovirus-K transmembrane envelope, Rec and Np9 proteins in melanomas and melanoma cell lines.** *Melanoma Res* 2006, **16:**223-234.

153. Hahn S, Ugurel S, Hanschmann KM, Strobel H, Tondera C, Schadendorf D, Lower J, Lower R: **Serological response to human endogenous retrovirus K in melanoma patients correlates with survival probability.** *AIDS Res Hum Retroviruses* 2008, **24:**717-723.

154. Muster T, Waltenberger A, Grassauer A, Hirschl S, Caucig P, Romirer I, Fodinger D, Seppele H, Schanab O, Magin-Lachmann C, et al: **An endogenous retrovirus derived from human melanoma cells.** *Cancer Res* 2003, **63:**8735-8741.

155. Contreras-Galindo R, Kaplan MH, Markovitz DM, Lorenzo E, Yamamura Y: **Detection of HERV-K(HML-2) viral RNA in plasma of HIV type 1-infected individuals.** *AIDS Res Hum Retroviruses* 2006, **22:**979-984.

156. Contreras-Galindo R, Lopez P, Velez R, Yamamura Y: **HIV-1 infection increases the expression of human endogenous retroviruses type K (HERV-K) in vitro.** *AIDS Res Hum Retroviruses* 2007, **23:**116-122.

157. Garrison KE, Jones RB, Meiklejohn DA, Anwar N, Ndhlovu LC, Chapman JM, Erickson AL, Agrawal A, Spotts G, Hecht FM, et al: **T cell responses to human endogenous retroviruses in HIV-1 infection.** *PLoS Pathog* 2007, **3:**e165.

158. Seifarth W, Frank O, Zeilfelder U, Spiess B, Greenwood AD, Hehlmann R, Leib-Mosch C: **Comprehensive analysis of human endogenous retrovirus transcriptional activity in human tissues with a retrovirus-specific microarray.** *J Virol* 2005, **79:**341-352.

159. Faff O, Murray AB, Schmidt J, Leib-Mosch C, Erfle V, Hehlmann R: **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):**1087-1097.

160. Seifarth W, Baust C, Murr A, Skladny H, Krieg-Schneider F, Blusch J, Werner T, Hehlmann R, Leib-Mosch C: **Proviral structure, chromosomal location, and expression of HERV-K-T47D, a novel human endogenous retrovirus derived from T47D particles.** *J Virol* 1998, **72:**8384-8391.

161. Ejthadi HD, Martin JH, Junying J, Roden DA, Lahiri M, Warren P, Murray PG, Nelson PN: **A novel multiplex RT-PCR system detects human endogenous retrovirus-K in breast cancer.** *Arch Virol* 2005, **150:**177-184.

162. Golan M, Hizi A, Resau JH, Yaal-Hahoshen N, Reichman H, Keydar I, Tsarfaty I: **Human endogenous retrovirus (HERV-K) reverse transcriptase as a breast cancer prognostic marker.** *Neoplasia* 2008, **10:**521-533.

163. Galli UM, Sauter M, Lecher B, Maurer S, Herbst H, Roemer K, Mueller-Lantzsch N: **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:**3223-3228.

164. Yi JM, Kim HM, Kim HS: **Expression of the human endogenous retrovirus HERV-W family in various human tissues and cancer cells.** *J Gen Virol* 2004, **85:**1203-1210.

165. Huang WJ, Liu ZC, Wei W, Wang GH, Wu JG, Zhu F: **Human endogenous retroviral pol RNA and protein detected and identified in the blood of individuals with schizophrenia.** *Schizophr Res* 2006, **83:**193-199.

166. Karlsson H, Bachmann S, Schroder J, McArthur J, Torrey EF, Yolken RH: **Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia.** *Proc Natl Acad Sci* 2001, **98:**4634-4639.

167. Otowa T, Tochigi M, Rogers M, Umekage T, Kato N, Sasaki T: **Insertional polymorphism of endogenous retrovirus HERV-K115 in schizophrenia.** *Neurosci Lett* 2006, **408:**226-229.

168. Liu SL, Miller AD: **Oncogenic transformation by the jaagsiekte sheep retrovirus envelope protein.** *Oncogene* 2007, **26:**789-801.

169. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.

170. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, et al: **The UCSC Genome Browser database: update 2010.** *Nucleic Acids Res*, **38:**D613-619.

171. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.

172. Hall TA: **BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.** *Nucleic Acids Symp Ser* 1999**:**95−98.

173. Barbulescu M, Turner G, Seaman MI, Deinard AS, Kidd KK, Lenz J: **Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans.** *Curr Biol* 1999, **9:**861-868.

174. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132:**365-386.

175. Torrey EF, Webster M, Knable M, Johnston N, Yolken RH: **The Stanley Foundation brain collection and Neuropathology Consortium.** *Schizophrenia Research* 2000, **44:**151-155.

176. Kumar S, Tamura K, Jakobsen IB, Nei M: **MEGA2: molecular evolutionary genetics analysis software.** *Bioinformatics* 2001, **17:**1244-1245.

177. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19:**1572-1574.

178. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754-755.

179. Calle EE, Rodriguez C, Jacobs EJ, Almon ML, Chao A, McCullough ML, Feigelson HS, Thun MJ: **The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics.** *Cancer* 2002, **94:**500-511.

180. Stoye J, Frankel W, Coffin J: **DNA hybridization in dried gels with fragmented probes:an improvement over blotting techniques.** *Technique* 1991, **3:**123-128.

181. Flockerzi A, Maydt J, Frank O, Ruggieri A, Maldener E, Seifarth W, Medstrand P, Lengauer T, Meyerhans A, Leib-Mosch C, et al: **Expression pattern analysis of transcribed HERV sequences is complicated by ex vivo recombination.** *Retrovirology* 2007, **4:**39.

182. Jha AR, Pillai SK, York VA, Sharp ER, Storm EC, Wachter DJ, Martin JN, Deeks SG, Rosenberg MG, Nixon DF, Garrison KE: **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:**2617-2626.

183. Bannert N, Kurth R: **Retroelements and the human genome: new perspectives on an old relation.** *Proc Natl Acad Sci U S A* 2004, **101 Suppl 2:**14572-14579.

184. Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, Lenz J: **Insertional polymorphisms of full-length endogenous retroviruses in humans.** *Curr Biol* 2001, **11:**1531-1535.

185. Reus K, Mayer J, Sauter M, Zischler H, Muller-Lantzsch N, Meese E: **HERV-K(OLD): ancestor sequences of the human endogenous retrovirus family HERV-K(HML-2).** *J Virol* 2001, **75:**8917-8926.

186. Costas J: **Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length proviral genomes.** *J Mol Evol* 2001, **53:**237-243.

187. Tonjes RR, Czauderna F, Kurth R: **Genome-wide screening, cloning, chromosomal assignment, and expression of full-length human endogenous retrovirus type K.** *J Virol* 1999, **73:**9187-9195.

188. Sugimoto J, Matsuura N, Kinjo Y, Takasu N, Oda T, Jinno Y: **Transcriptionally active HERV-K genes: identification, isolation, and chromosomal mapping.** *Genomics* 2001, **72:**137-144.

189. Lebedev YB, Belonovitch OS, Zybrova NV, Khil PP, Kurdyukov SG, Vinogradova TV, Hunsmann G, Sverdlov ED: **Differences in HERV-K LTR insertions in orthologous loci of humans and great apes.** *Gene* 2000, **247:**265-277.

190. Lower R, Boller K, Hasenmaier B, Korbmacher C, Muller-Lantzsch N, Lower J, Kurth R: **Identification of human endogenous retroviruses with complex mRNA expression and particle formation.** *Proc Natl Acad Sci U S A* 1993, **90:**4480-4484.

191. Blomberg J, Ushameckis D, Jern P: **Evolutionary aspects of human endogenous retroviral sequences (HERVs) and disease.** In *Retroviruses and Primate Genome Evolution.* Edited by Sverdlov ED. Georgetown, TX: Eurekah.com, Landes Bioscience; 2005: 204-239

192. Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE: **Characterization of six human disease-associated inversion polymorphisms.** *Hum Mol Genet* 2009, **18:**2555-2566.

193. Ji X, Zhao S: **DA and Xiao-two giant and composite LTR-retrotransposon-like elements identified in the human genome.** *Genomics* 2007**:**249-258.

194. Li X, Slife J, Patel N, Zhao S: **Stepwise evolution of two giant composite LTR0retrotransposon-like elements DA and Xiao.** *BMC Evolutionary Biology* 2009, **9**.

195. Newman T, Trask BJ: **Complex evolution of 7E olfactory receptor genes in segmental duplications.** *Genome Res* 2003, **13**.

196. She X, Liu G, Ventura M, Zhao S, Misceo D, Roberto R, Cardone MF, Rocchi M, Green ED, Archidiacano N, Eichler EE: **A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications.** *Genome Res* 2006, **16:**576-583.

197. Sugawara H, Harada N, Ida T, Ishida T, Ledbetter DH, Yoshiura K, Ohta T, Kishino T, Niikawa N, Matsumoto N: **Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23.** *Genomics* 2003, **82:**238-244.

198. Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, Neumann T, Ohashi H, Voullaire L, Larizza D, Giorda R, et al: **Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements.** *Am J Hum Genet* 2001, **68:**874-883.

199. Taudien S, Galgoczy P, Huse K, Reichwald K, Schilhabel M, Szafranski K, Shimizu A, Asakawa S, Frankish A, Loncarevic IF, et al: **Polymorphic segmental duplications at 8p23.1 challenge the determination of individual defensin gene repertoires and the assembly of a contiguous human reference sequence.** *BMC Genomics* 2004, **5:**92.

200. Bosch N, Morell M, Ponsa I, Mercader JM, Armengol L, Estivill X: **Nucleotide, cytogenetic and expression impact of the human chromosome 8p23.1 inversion polymorphism.** *PloS one* 2009, **4:**9.

201. Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE: **Ancestral reconstruction of segmental duplciations reveals punctated cores of human genome evolution.** *Nature Genetics* 2007, **39**.

202. Murphy WJ, Larkin DM, van der Wind AE, Bourque JE, Tesler G, Auvil L, Beever JE, Chowdhary BP, Galibert F, Gatzke L, et al: **Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps.** *Science* 2002, **309**.

203. Page RDM, Holmes EC: **Genes, organisation, function, and evolution.** In *Molecular Evolution: A Phylogenetic Approach.* Oford, UK: Blackwell Science, Ltd.; 1998: 72-88

204. Krzysztalowska-Wawrzyniak M, Ostanek M, Clark J, Binczak-Kuleta A, Ostanek L, Kaczmarczyk M, Loniewska B, Wyrwicz LS, Brzosko M, Ciechanowicz A: **The distribution of human endogenous retrovirus K-113 in health and autoimmune diseases in Poland.** *Rheumatology (Oxford)*, **50:**1310-1314.

205. Burmeister T, Ebert AD, Pritze W, Loddenkemper C, Schwartz S, Thiel E: **Insertional polymorphisms of endogenous HERV-K113 and HERV-K115 retroviruses in breast cancer patients and age-matched controls.** *AIDS Res Hum Retroviruses* 2004, **20:**1223-1229.

206. Frankel WN, Stoye JP, Taylor BA, Coffin JM: **A linkage map of endogenous murine leukemia proviruses.** *Genetics* 1990, **124:**221-236.

207. Larsson LI, Bjerregaard B, Wulf-Andersen L, Talts JF: **Syncytin and cancer cell fusions.** *Scientific World Journal* 2007, **7:**1193-1197.

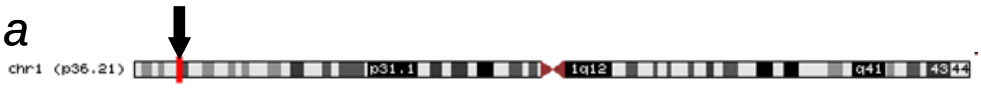
208. Strick R, Ackermann S, Langbein M, Swiatek J, Schubert SW, Hashemolhosseini S, Koscheck T, Fasching PA, Schild RL, Beckmann MW, Strissel PL: **Proliferation and cell-cell fusion of endometrial carcinoma are induced by the human endogenous retroviral Syncytin-1 and regulated by TGF-beta.** *J Mol Med* 2007, **85:**23-38.

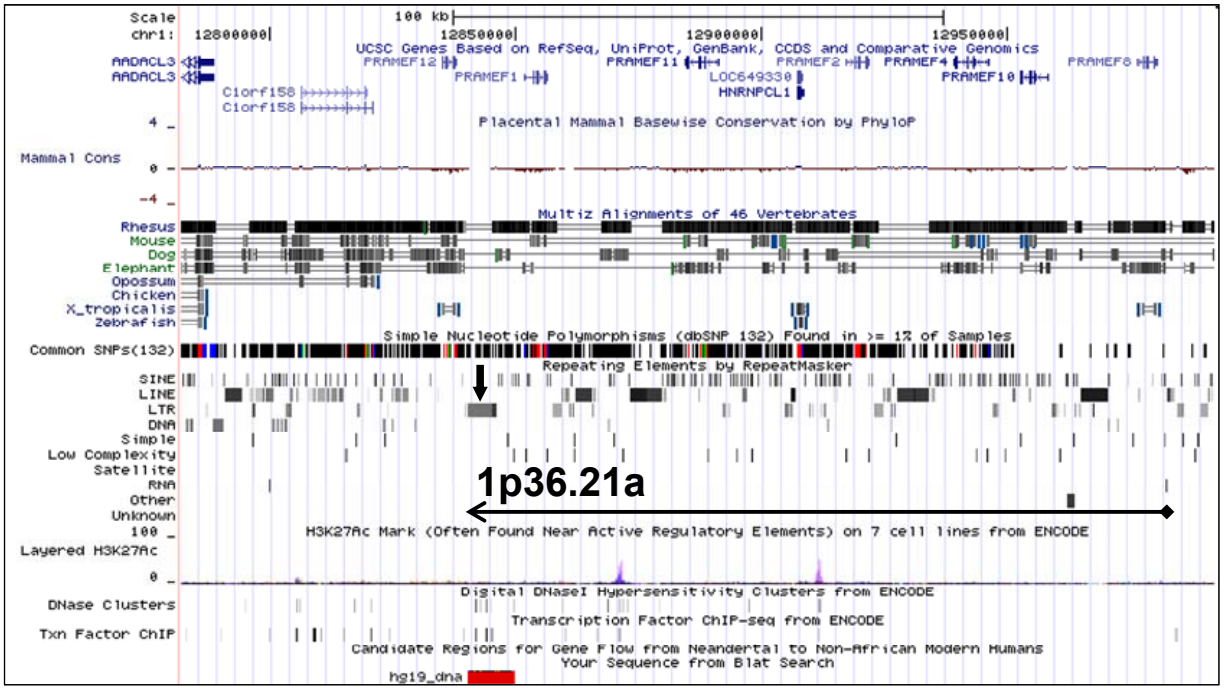
209. Moyes DL, Goris A, Ban M, Compston A, Griffiths DJ, Sawcer S, Venables PJ: **HERV-K113 is not associated with multiple sclerosis in a large family-based study.** *AIDS Res Hum Retroviruses* 2008, **24:**363-365.

210. Ruprecht K, Mayer J, Sauter M, Roemer K, Mueller-Lantzsch N: **Endogenous retroviruses and cancer.** *Cell Mol Life Sci* 2008, **65:**3366-3382.

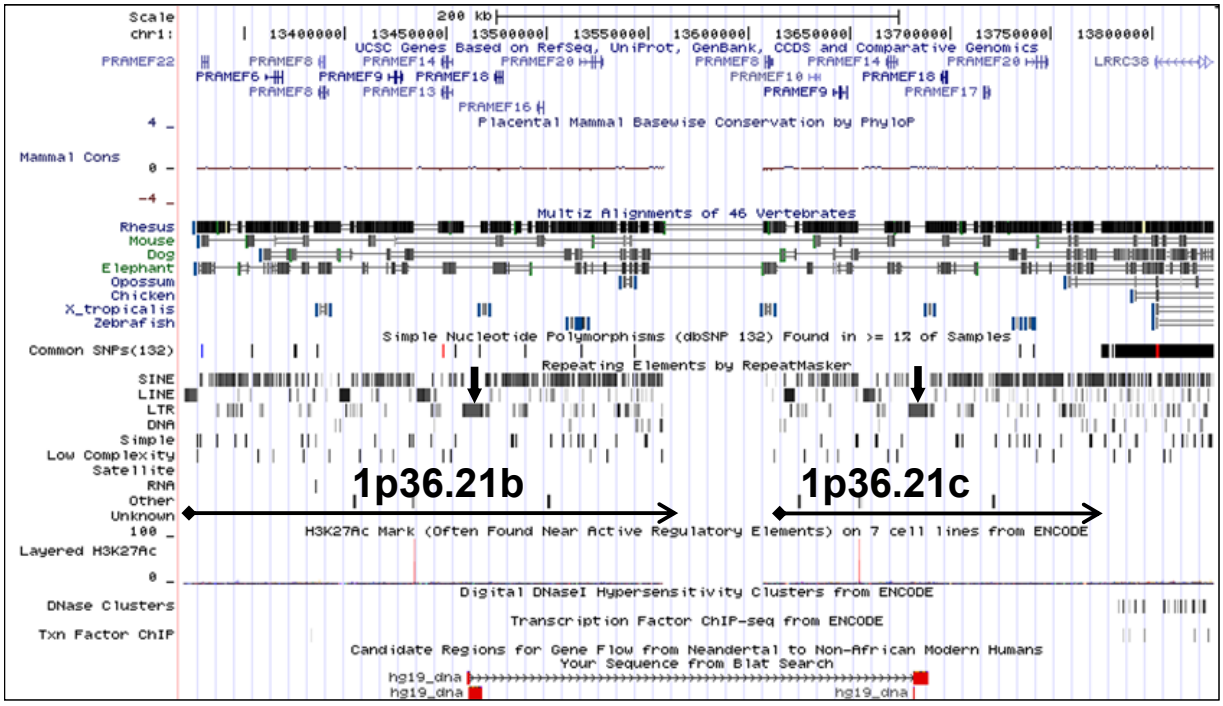
APPENDIX A

1p36.21 a

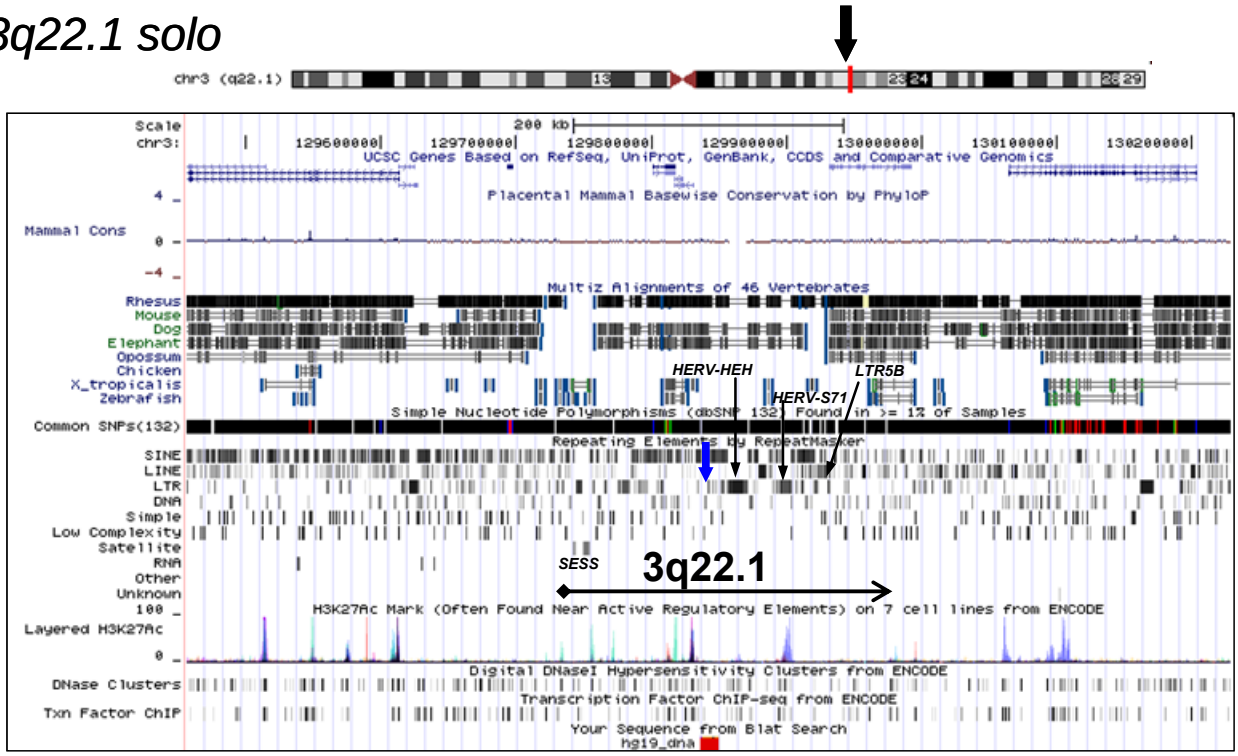




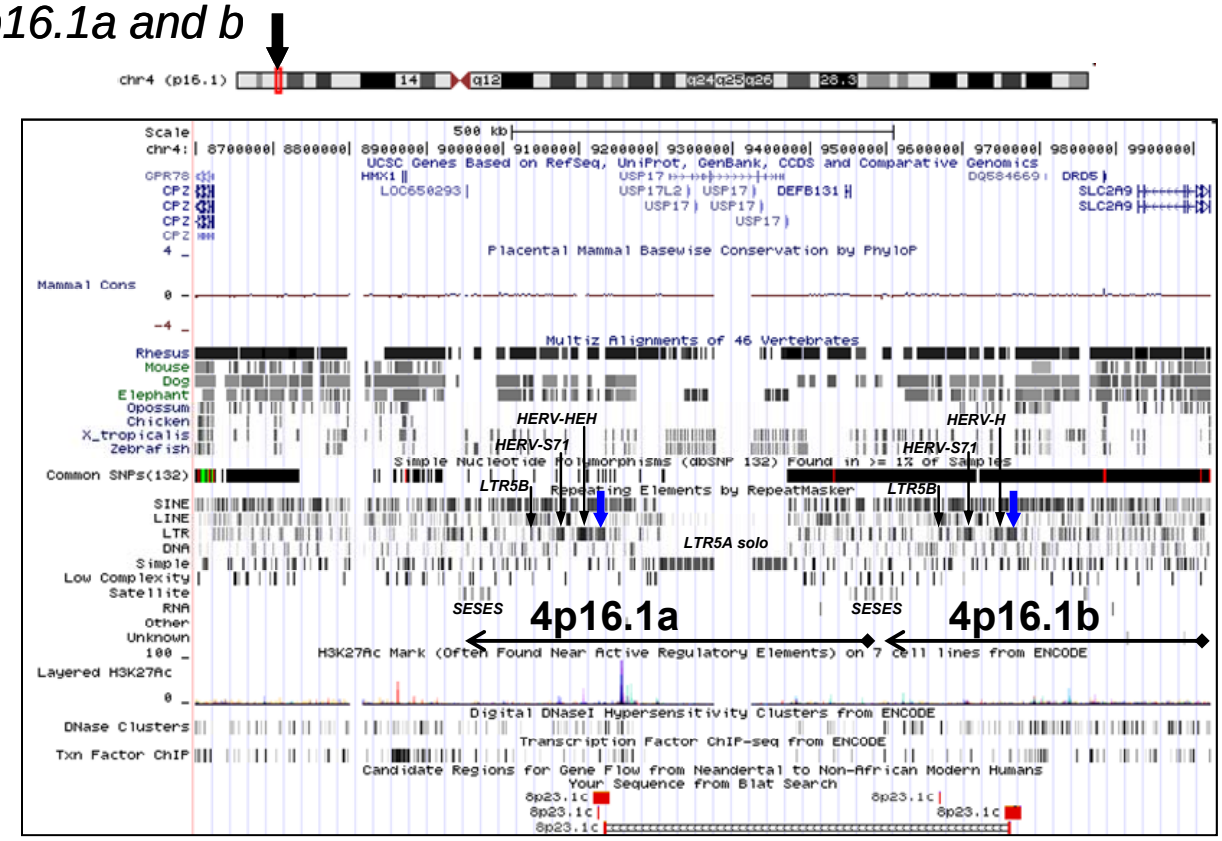
1p36.21 b and c



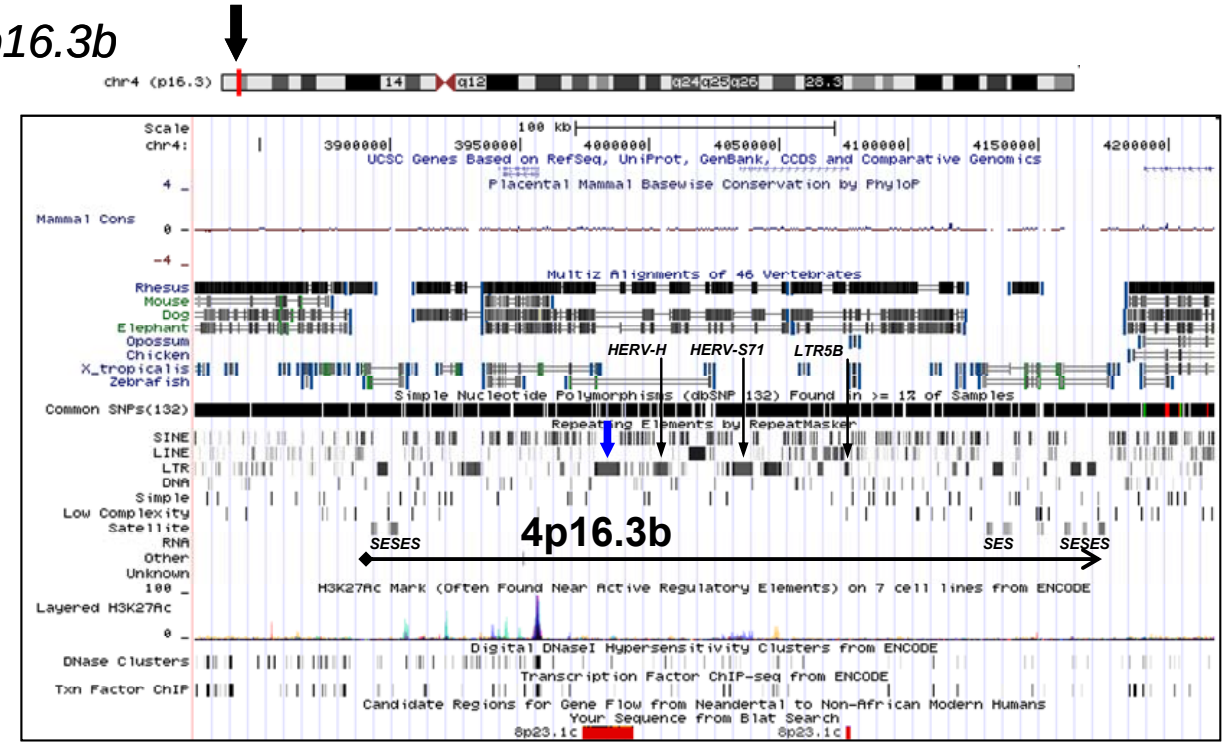
3q22.1 solo


chr3 (q22.1)
Scale
chr3:
200 kb
129600000
129700000
129800000
129900000
130000000
130100000
130200000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
HERV-HEH
HERV-S71
LTR5B
Simple Nucleotide Polymorphisms (dbSNP 132) Found in >= 1% of Samples
Common SNPs(132)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
SESS
3q22.1
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Your Sequence from Blat Search
hg19_dna

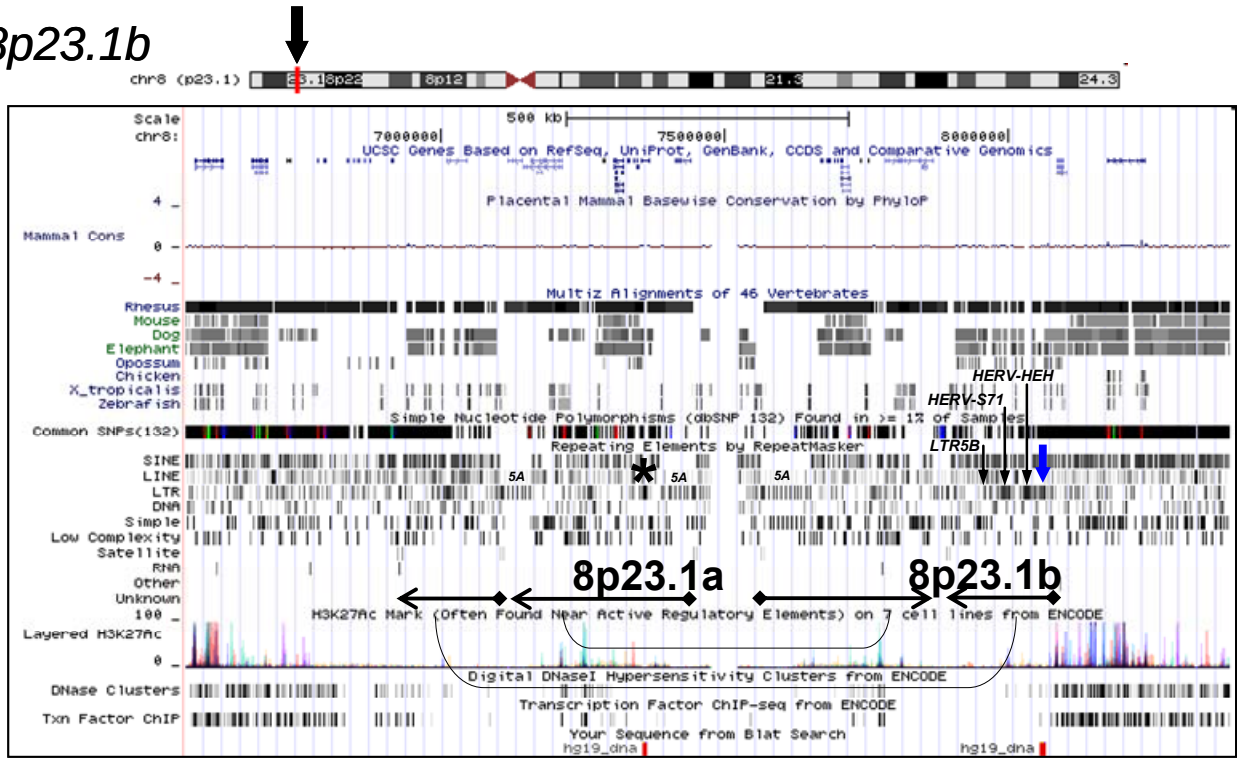
4p16.1a and b



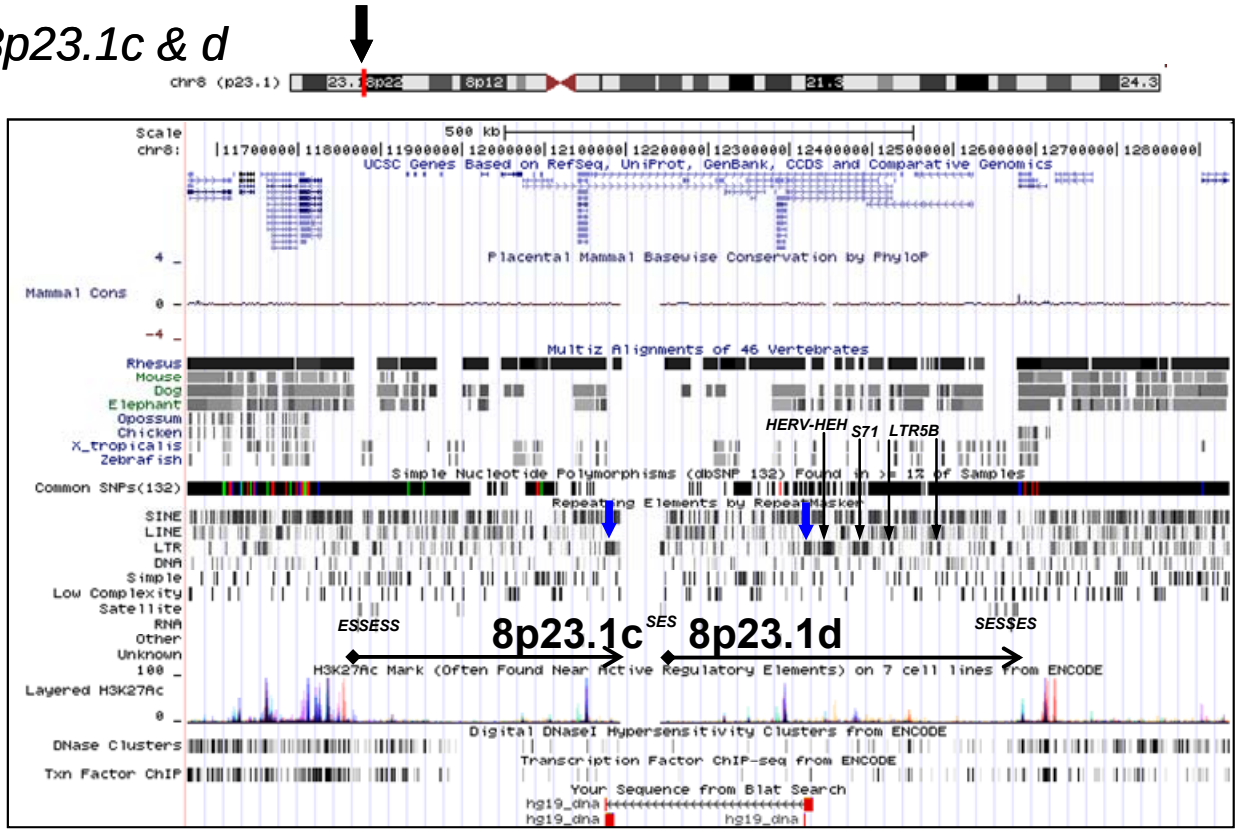
4p16.3b



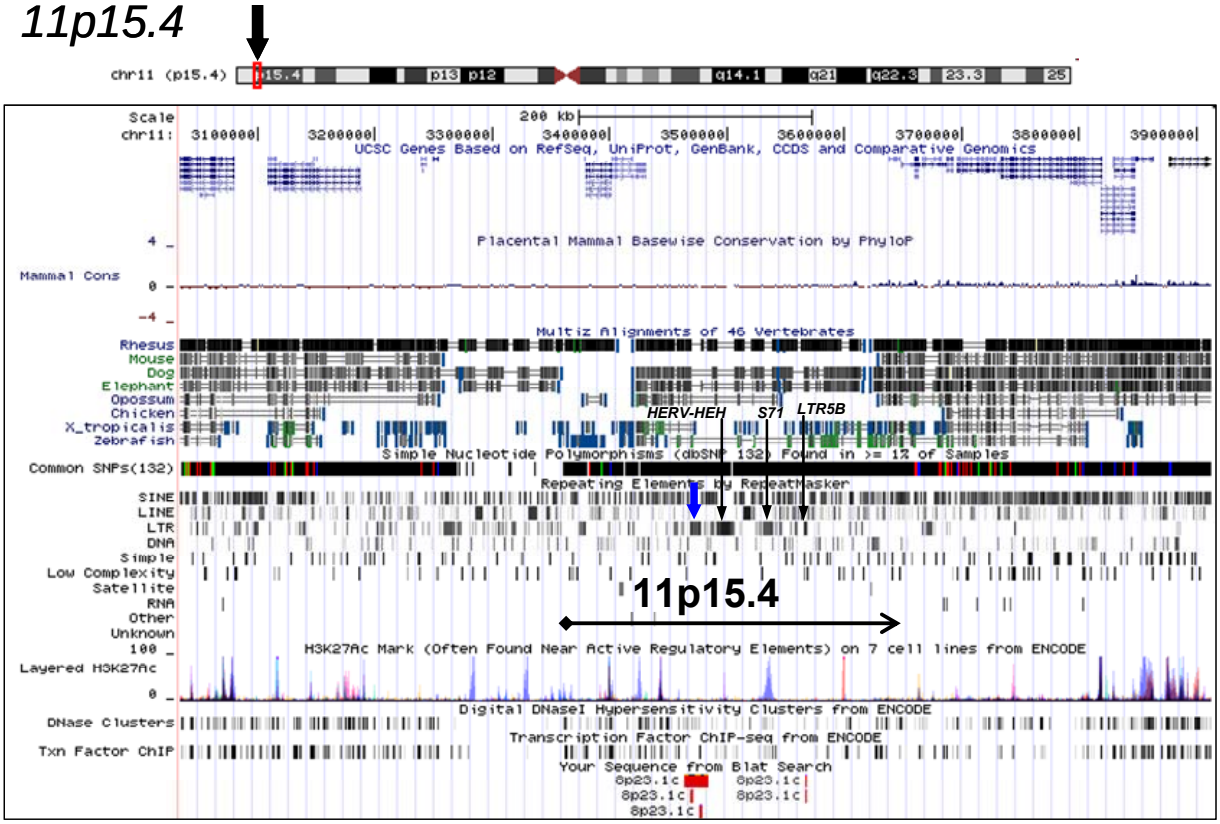
8p23.1b



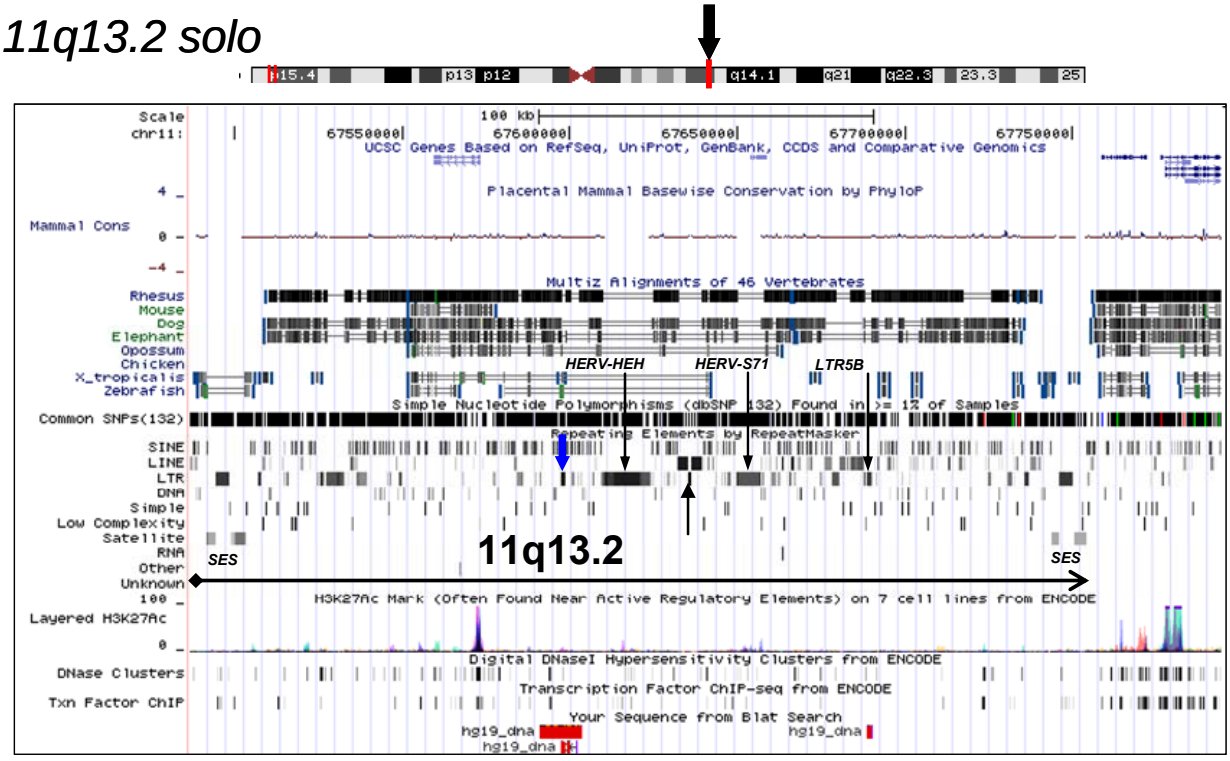
8p23.1c & d



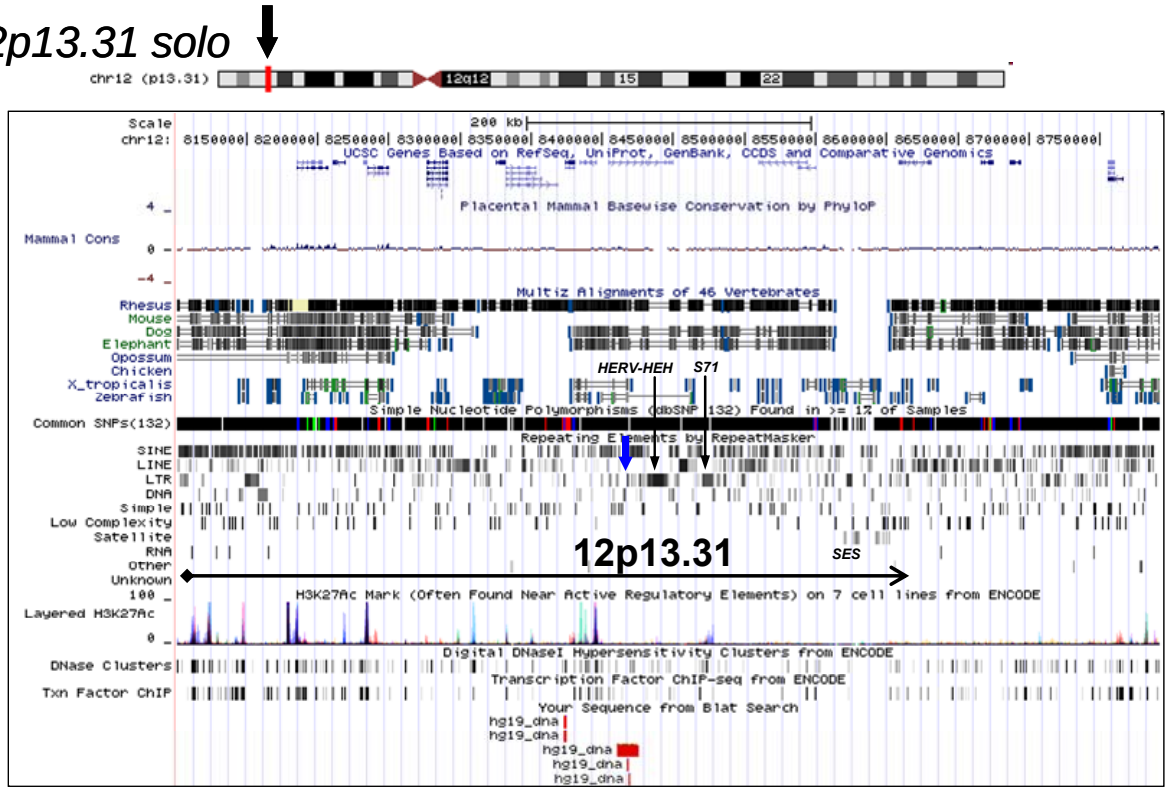
11p15.4



11q13.2 solo



12p13.31 solo



Xq28a and b