2016

Paleovirological Analyses of Endogenous Retroviruses and Host Innate Immune Effectors

Daniel Blanco Melo

PALEOVIROLOGICAL ANALYSES OF ENDOGENOUS RETROVIRUSES AND HOST INNATE IMMUNE EFFECTORS

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Daniel Blanco Melo

June 2016

PALEOVIROLOGICAL ANALYSES OF ENDOGENOUS RETROVIRUSES AND HOST INNATE IMMUNE EFFECTORS

Daniel Blanco Melo, Ph.D.

The Rockefeller University 2016

About 8 and 10 percent of the human and mouse genomes, respectively, are comprised of sequences of retroviral origin. Occasional infection of germ line cells can lead to integrated retroviral genomes being vertically inherited as host alleles. During thousands to millions of years, some of these sequences acquired inactivating mutations and were fixed in ancestral populations by genetic drift, while others became fixed by providing an evolutionary advantage to the host. Those inherited proviruses are termed endogenous retroviruses (ERVs) and have been identified in a variety of animal species representing an extensive viral "fossil" record of past retroviral infections. With the advent of whole genome sequencing projects and high throughput sequencing platforms, it became evident the wide diversity and the important role that these sequences have had in the evolution of their hosts. In the present study we developed a computational framework to identify ERVs in primate and murine genomes. The results of these genome screenings were used to identify suitable candidate sequences in which to perform paleovirological analyses that lead to the successful reconstruction of two ancient retroviruses.

MuERV-L is an *env*-deficient highly abundant mouse specific ERV that has undergone two amplification bursts, being the more recent and prolific ~2 million

years ago (MYA), probably through entirely intracellular mechanisms. MuERV-L is transcriptionally active at the two-cell stage of the mouse embryo and recent studies have implicated the co-option of its LTR as a promoter for totipotency genes. In the present work, we describe the analysis and reconstruction of an infectious ancestral MuERV-L (ancML) sequence through paleovirological analyses of MuERV-L elements in the mouse genome. The resulting ancML sequence was infectious in CHO cells and its replication was dependent on reverse transcription. We found that IFN-α could reduce ancML replication by ~20 fold. Additionally, we found that the expression of mouse APOBEC3 was able to restrict the replication of ancML. However, inspection of endogenous MuERV-L sequences suggested that the impact of APOBEC3 mediated hypermutation on MuERV-L evolution was limited. We discussed the possibility that type I IFN responses (maybe through restriction factors) might inhibit MuERV-L replication at the two-cell stage of the mouse embryo and have kept MuERV-L copy numbers under control.

Although no extant human gammaretroviruses have been identified, HERV-T is a low copy primate ERV lineage that is closely related to the gammaretrovirus genus. Through phylogenetic and genomic analysis of HERV-T insertions we defined three distinct lineages. Two lineages (HERV-T1 and HERV-T2) entered the primate germline after the Old World monkey-ape split about ~32-30 MYA, whereas the other (HERV-T3) entered before this divergence ~40 MYA. Phylogenetic analysis of complete (LTR-*gag*-*pol*-*env*-LTR) proviral sequences

showed that HERV-T2 was subjected to APOBEC3 mediated hypermutation, and subsequently expanded in apes, most likely through retrotransposon-like mechanisms. Phylogenetic and statistical analysis of HERV-T3 proviruses allowed us to estimate the sequence of their ~32 MY old ancestor, revealing that its unusually long leader sequence encoded a 855-nucleotide ORF separated from gag by 36 nucleotides. This *pre-gag* ORF of unknown function putatively codes for a protein that includes a transmembrane domain. Additional analysis of the HERV-T3 ancestral sequence allowed us to reconstruct the corresponding *env* sequence (ancHTenv). We found that a modern gammaretrovirus (MLV) could be pseudotyped with ancHTenv enabling it to infect a wide variety of primate cell lines with titers that are similar to MLV particles carrying the amphotropic MLV envelope. A single HERV-T proviral insertion in the genome of all great apes contains an *env* gene with full coding potential. Proteins encoded by the extant human HERV-T envelope gene (HsaHTenv) and one estimated to be encoded by the hominid ancestor were not able to generate infectious MLV pseudotyped particles, probably because HsaHTenv is not correctly processed into its mature and functional form. Statistical and phylogenetic analyses indicate that the *env* gene in this locus is evolving slower than the rest of the proviral sequences, and that selective pressures have acted on this locus to conserve its envelope sequence. Remarkably, we found that expression of the HsaHTenv was able to specifically block infection by MLV particles pseudotyped with the ancHTenv, but not particles pseudotyped with the amphotropic MLV

envelope. Additionally, we identified MOT1 as the receptor used by ancHTenv. Further experiments are needed in order to test the hypothesis that HsaHTenv served as a restriction factor through interference with the receptor once used by HERV-T.

As paleovirology also studies the evolution of the host defense mechanisms that have been shaped by past retroviral infections, we investigated the origins and evolution of *tetherin*, an orphan antiviral protein with no known homologs. We found that *tetherin* function is encoded by genes that exhibit no sequence homology and share only a common architecture and location in modern jawed vertebrate genomes, indicating an origin of ~450 MYA. Moreover, *tetherin* is part of a cluster of three potential sister genes that includes *pv1* and a putative gene of unknown function, here referred as *tm-cc(at)*, which encode proteins of similar architecture. Some variants of these proteins exhibit antiviral activity while others can be endowed with antiviral activity following a simple modification. Only in a slowly evolving species (coelacanths) does Tetherin exhibit homology to TM-CC(aT). We suggest that neofunctionalization, drift and positive selection drove a near complete loss of sequence similarity among modern *tetherin* genes, and between *tetherin* and its sister genes. Scenarios by which this orphan gene may have arisen and evolved exemplify how protein modularity, evolvability and robustness can create new functions and preserve them, despite sequence divergence due to genetic conflict with past and present viruses.

Quisiera dedicar este trabajo a todos los miembros de las familias Blanco y Melo. A mis abuelos, mis tíos y mis primos, vivos y ausentes, con los que he compartido muy gratas experiencias y que significan para mi mucho mas que un familiar.

Particularmente quisiera dedicar la culminación de mis estudios de doctorado a mis padres, Héctor Blanco Villasuso y a Aurora Melo Patiño, que con su constante apoyo a todas mis locuras han forjado el ser humano que soy hoy. Su cariño y amistad me han acompañado por momentos muy felices, pero también por momentos complicados, los cuales solo contribuyeron a fortalecer nuestros lazos. No hay un solo día que no agradezca por tenerlos en mi vida. ¡Los quiero mucho! Muchas, muchas gracias.

Acknowledgments

I would like to thank Dr. Paul D. Bieniasz for his guidance during my Ph.D. studies. His knowledge and his incredible ability to infer the causes of biological phenomena certainly were instrumental to bring three distinct projects to satisfactory completion. I am particularly thankful for his patience and understanding over the constant struggle to decipher the origins of *tetherin*.

I would like to thank all the current and past members of the Bieniasz lab for our many improvised chats and shared drinks, as well as their patience during my art history lessons / lab meetings. Especially I want to thank Sebla Kutluay, Matt McNatt, Rachel Liberatore, Melissa Kane, Theodora Hatziioannou, Steven Soll, Trinity Zang and Fengwen Zhang for their advice, help and understanding over my constant questions, favors and reagent requirements. I would like to thank Dr. Robert Gifford for his guidance in the exciting field of endogenous retroviruses. I am also particularly thankful to Dr. Siddarth Venkatesh for our pleasant and fructiferous collaboration.

I especially would like to thank Dr. Charles Rice for his support and much useful advice starting before my acceptance at Rockefeller, and continuing all through my years in this institution. I want to thank also Dr. David Ho for his advice as a member of my Faculty Advisory Committee and to Dr. Welkin Johnson for accepting my invitation to participate as an external examiner.

I would like to thank the Dean Sidney Strickland, Emily Harms and the faculty members of the admissions committee for giving me this exceptional opportunity. I am also thankful to Cris Rosario, Marta Delgado, Kristen Cullen and Stephanie Fernandez at the Dean's office for all their great help in a variety of student and personal life aspects.

The completion of my Ph.D. work could not have been possible without the indirect involvement of all my friends in New York City (NYC) that supported all my craziness and Mexican eccentricities. I am specially thankful to Sid Venkatesh, Julia Bitzegeio, Matt McNatt, Nina Gnädig, Tonya Kueck, Ashley York, Fabian Schmidt (the scientists), Michael Wheelock, Irit Shachrai, Gianna Triller (the Rockies), Ricardo Rocha, Mariana Peña, Yde Vázquez (the Mexicans), Emi Taipi, Linda Molla and Olta Molla (the Albanians) for their wonderful friendship and for reminding me of the Mexican warmth in these far north lands. I am particularly very thankful to Linda Molla for her affection and kindness. Të dua shume zemra.

Disclaimer

Dr. Siddarth Venkatesh contributed equally to our investigations into the evolutionary origins of *tetherin* and conducted the initial experiments concerning the antiviral activity of modified TM-CC proteins described in Chapter VI. Portions of chapters I, VI and VII are adapted from manuscripts that were co-written with Dr. Paul D. Bieniasz and Dr. Siddarth Venkatesh. Additionally, portions of chapter III are adapted from a manuscript that was written in collaboration with Dr. Robert Gifford.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

-sssDNA	:	Minus-stand strong stop DNA
+sssDNA	:	Plus-stand strong stop DNA
aa	:	Amino acid
AGS	:	Aicardi-Goutières syndrome
ALV	:	Avian leukemia virus
ancHsaHTenv	:	Recent Common Ancestor Of HsaHTenv
ancHsaHTenv-FurinFix	:	ancHsaHTenv with a functional furin cleavage site
ancHTenv	:	~32 my ancestral HERV-T envelope
ancHTenv-FurinMut	:	ancHTenv with a non-functional furin cleavage site
ancML	:	~2 my ancestral MuERV-L
ancML-RTmut	:	ancML mutant in the RT active site
APOBEC3	:	Apolipoprotein B editing catalytic subunit-like 3
AZT	:	Azidothymidine
BLAST	:	Basic local alignment search tool
bp	:	Base pairs
C-terminus	:	Carboxyl terminus
CA	:	Capsid protein
CAT-1	:	Cationic amino acid transporter 1
CC	:	Coiled-coil
cDNA	:	Complimentary DNA
CERV	:	Chimpanzee ERV
cfu	:	Colony forming units
CHO	:	Chinese hamster ovary
CMV	:	Cytomegalovirus
CT	:	Cytoplasmic tail

CTR1	:	Copper transport protein 1
cypA	:	Cyclophillin A
dA	:	2’-deoxyadenine
DAPI	:	4',6-diamidino-2-phenylindole
dC	:	2’-deoxycytidine
dG	:	2’-deoxyguanine
DIGS	:	Database integrated genome screening
dN	:	Ratio of non-synonymous substitutions over non-synonymous sites
dNTPs	:	Deoxynucleoside triphosphates
dpi	:	Days post infection
dpt	:	Days post transfection
dS	:	Ratio of synonymous substitutions over synonymous sites
dsDNA	:	Double stranded DNA
dT	:	2’-deoxytimidine
dU	:	2’-deoxyuridine
dUTP	:	2'-deoxyuridine 5'-triphosphate
enJSRV	:	Endogenous JSRV
ER	:	Endoplasmic reticulum
ERV	:	Endogenous retrovirus
EVE	:	Endogenous viral elements
FACS	:	Fluorescence-activated cell sorting
FeLV	:	Feline leukemia virus
GaLV	:	Gibbon ape leukemia virus
gDNA	:	Genomic DNA
GFP	:	Green fluorescent protein
GlycoGag	:	MLV glycosylated gag
GPI	:	Glycophosphatidylinositol anchor

GRC	:	Genome reference consortium
GTR	:	General time reversible (evolutionary model)
HA-tag	:	Human influenza hemagglutinin derived tag
HERV	:	Human ERV
HERV-L	:	Human endogenous retrovirus - L
HERV-T	:	Human endogenous retrovirus - T
HIV	:	Human immunodeficiency virus
HsaHTenv	:	HERV-T envelope with full coding capacity present in the human genome
HsaHTenv-FurinFix	:	HsaHTenv with a functional furin cleavage site
HTLV	:	Human t-cell lymphotropic virus
HTpG	:	HERV-T pre-gag protein
IFN	:	Interferon
IN	:	Integrase
JSRV	:	Jaagsiekte sheep retrovirus
KB	:	Kilo bases
kDa	:	Kilo daltons
KoRV	:	Koala retrovirus
lncRNA	:	Long ncRNA
LTR	:	Long terminal repeats
MA	:	Matrix
mA3	:	Mouse APOBEC3 protein
MB	:	Mega bases
ML	:	Maximum likelihood
MLV	:	Murine leukemia virus
MLV-A	:	MLV amphotropic envelope
MLV-E	:	MLV ecotropic envelope
MLV-P	:	MLV polytropic envelope
MLV-X	:	MLV xenotropic envelope

MMTV	:	Mouse mammary tumor virus
MOI	:	Multiplicity of infection
MOT1	:	Monocarboxylate transporter 1
mRNA	:	Messenger RNA
MSA	:	Multiple sequence alignment
MuERV-L	:	Murine endogenous retrovirus - L
MuERV-Lref	:	MuERV-L reference sequence from Benit et.al. 1997
MY	:	Million years
MYA	:	Million years ago
N-terminus	:	Amino terminus
NC	:	Nucleocapsid
ncRNA	:	Non coding RNA
NEO	:	Neomycin resistance gene
nt	:	Nucleotide
ORF	:	Open reading frame
PBS	:	Primer binding site
PCR	:	Polymerase chain reaction
PIC	:	Pre-integration complex
PPT	:	Polypurine tract
PR	:	Protease
pSIV	:	Prosimian immunodeficiency virus
RELIK	:	Rabbit endogenous lentivirus type K
RFP	:	Red fluorescent protein
RNA-Seq	:	RNA sequencing
RNAP II	:	RNA polymerase II
RT	:	Reverse transcriptase
SAMHD1	:	SAM domain and HD domain-containing protein 1
SFV	:	Simian foamy virus

soloLTR	:	Single LTR left after excision of the internal sequences.
SU	:	Surface protein
SV40	:	Simian virus 40
TM	:	Tramsmembrane protein or domain
tm-cc(at)	:	Tm and cc encoding gene adjacent to *tetherin*
TRIM	:	Tripartite motif
tRNA	:	Transfer RNA
UTR	:	Un-translated region
vgRNA	:	Viral genomic RNA
VSV-g	:	Vesicular stomatitis virus glycoprotein
WGS	:	Whole genome shotgun (assembly)
WMSV	:	Woolly monkey sarcoma virus
WSDV	:	Walleye dermal sarcoma virus
ZGA	:	Zygote genome activation
Ψ	:	Packaging signal

Chapter I. Introduction

Viruses are the most abundant biological entities on the planet (Edwards and Rohwer, 2005) accounting for approximately 94% of all nucleic-acid containing particles present in the oceans (Suttle, 2007). It is estimated that there are in total 10^{31} viruses on Earth (Breitbart and Rohwer, 2005), and have been found infecting organisms from each of the three domains of life (Koonin et al., 2006). Although there is still a debate as to whether viruses originated before or after the last universal cellular ancestor (Forterre, 2006; Holmes, 2011; Koonin et al., 2015), the fact is that viruses have existed for many millions of years. Unlike other organisms, viruses have not left a physical fossil record by which researchers can document the existence of an extinct lineage or follow their evolutionary path from an ancient specie into a modern one throughout millions of years (MY). This fact was generally accepted until integrated retroviral sequences were found in normal chicken and mouse embryos in the late 1960s (Aaronson et al., 1969; Weiss, 1967; Weiss, 2006). Since then, and with the advent of whole genome projects and high throughput sequencing platforms, the wide diversity of viral sequences present in animal genomes and the important role that these sequences have had in the evolution of their hosts became evident. Currently, integrated sequences (known as endogenized viral sequences) for almost all classes of viruses have been found in different eukaryotic genomes (Aiewsakun and Katzourakis, 2015). However, there is one group of viruses that outnumber all the rest, the retroviruses.

Retroviruses: structure, classification and replication.

The *Retroviridae* family is composed of positive-sense RNA enveloped viruses with two obligate and defining features: (i) the synthesis of a DNA intermediate from an RNA genome (a process termed reverse transcription), and (ii) the integration of their genetic information into the host genome (resulting in a structure termed provirus) (Coffin et al., 1997). This mandatory integration step creates a unique and intimate relationship between the virus and the host that has shaped the evolution of both entities, and it is the reason why the vast majority of endogenous viral elements (EVEs) in animal genomes are from retroviral origin.

Retroviruses can be classified into simple or complex depending on the gene set they encode. The minimal structure of the provirus of a simple retrovirus is depicted in (Figure 1.1A):

Long Terminal Repeats (LTR): Direct repeats found flanking the internal sequence of a provirus and it is subdivided into three regions. U3 is a sequence unique to the 3' end of the viral genome that promotes the transcription of the provirus by containing a RNA polymerase II (RNAP II) promoter and several binding sites for transcription factors. The R region is a short sequence repeated in both 5' and 3' ends of the viral genome. It is involved in strand transfer during reverse transcription (see below) and contains signals involved in the genomic/mRNA 3' end cleavage and polyadenylation. U5 is a sequence unique

to the 5' end of the viral genome and contains signals that promote transcriptional termination and polyadenylation (Bohnlein et al., 1989; Fields et al., 2001).

gag (stands for group-specific antigen): Encodes a polyprotein that is cleaved to produce the internal structural proteins of the virus (Figure 1.1A and C). The Gag polyprotein contains at least three proteins: matrix (MA), which is N-terminally myristoylated and targets particle assembly to the plasma membrane (possibly through interaction with host tRNAs) (Bieniasz, 2009; Kutluay et al., 2014); capsid (CA), which forms the internal virion core; and nucleocapsid (NC), which binds to viral genomic RNA and drives genome packaging into virions (Kutluay et al., 2014; Pedersen et al., 2011).

pol: Encodes non-structural proteins required for viral replication (Figure 1.1A and C). The Pol polyprotein is incorporated into viral particles joined to Gag and it contains: the reverse transcriptase (RT), involved in viral replication and shows RNA- and DNA-dependent DNA polymerase and RNase H activities (Champoux and Schultz, 2009); and integrase (IN), which mediates the integration of the viral DNA into the host genome (Fields et al., 2001; Pedersen et al., 2011). In some retroviruses *pol* also encodes for a third protein, the protease (PR), which drives the processing of the Gag and Gag-Pol polyproteins into their functional units during particle maturation (Sundquist and Krausslich, 2012). PR can also be fused to the 3' end of the *gag* gene or encoded by a separate ORF between *gag* and *pol* (Coffin et al., 1997) (Figure 1.1D).

env: Encodes the viral envelope, which interacts with cellular receptors leading to the fusion of the viral and cell membranes resulting in virus entry. The *env* gene is expressed from a different subgenomic mRNA that is joined to a 5' leader sequence by splicing, removing the *gag* and *pol* genes (Fields et al., 2001) (Figure 1.1A). An N-terminal short hydrophobic signal peptide leads the translocation of the Env precursor into the lumen of the endoplasmic reticulum (ER) (stopping at its transmembrane domain), where it is glycosylated, folded and oligomerized (Figure 1.1B). In order to drive membrane fusion, the oligomerized Env precursor needs further proteolytic processing into surface (SU) and transmembrane (TM) proteins (Figure 1.1A, B and C) (Coffin et al., 1997), mediated by the host protease furin in the Golgi apparatus (Hallenberger et al., 1992; Stein and Engleman, 1990). This cleavage allows the hydrophobic fusion peptide (located in the N-terminus of the TM protein) to mediate fusion of the cellular and viral membranes after the specific SU-receptor interaction (Pedersen et al., 2011) (Figure 1.1B).

Some retroviruses also include another gene encoding for a dUTPase in distinct locations. The product of this gene reduces the incorporation of dUTP into viral DNA, which could induce mutations during viral replication (Lerner et al., 1995). Moreover, all retroviruses contain additional signals that are necessary for reverse transcription, such as the primer binding site (PBS) and the polypurine tract (PPT) (see below), genome packaging such as the packaging signal (psi or Ψ) recognized by NC, and other regulatory elements (Figure 1,1A).

In addition to all these major components, complex retroviruses also encode for “accessory” genes (Figure 1.1D), which in general encode proteins that regulate and coordinate viral gene expression, RNA processing, inhibit host defenses and other secondary roles. Accessory genes can be located in different regions of the viral genome between distinct retroviruses and they are all expressed from spliced RNAs (Figure 1.1D).

Figure 1.1: Retroviral genome and particle structure.

(A) Structure of a simple provirus, viral genomic RNA/mRNA, spliced mRNA and main protein products. Refer to main text for gene and protein symbols. PBS: primer binding site (orange box). PPT: polypurine tract (clear blue box). Ψ: packaging signal. SD: splice donor site. SA: splice acceptor site. AAA_n: 3' poly(A) tail. 5' cap is shown as a dark blue half circle. Grey, cyan and purple boxes indicate U3, R and U5 regions of the LTR, respectively. **(B)** Organization and processing of the MLV Env glycoprotein. Amino acid positions of each processed segment is shown below the black line. Red and orange boxes indicate signal and fusion peptides, respectively. Host and viral protease cleavage events are shown as blue and white thunderbolts, respectively. **(C)** Structure of a retroviral particle indicating viral proteins described in (A). **(D)** Genome structures of simple and complex retroviruses. ORFs in the same frame are shown in the same line. Dashed lines indicate spliced introns. (A and C) adapted from Pedersen et al., 2006. (B and D) adapted from Coffin et al., 1997.

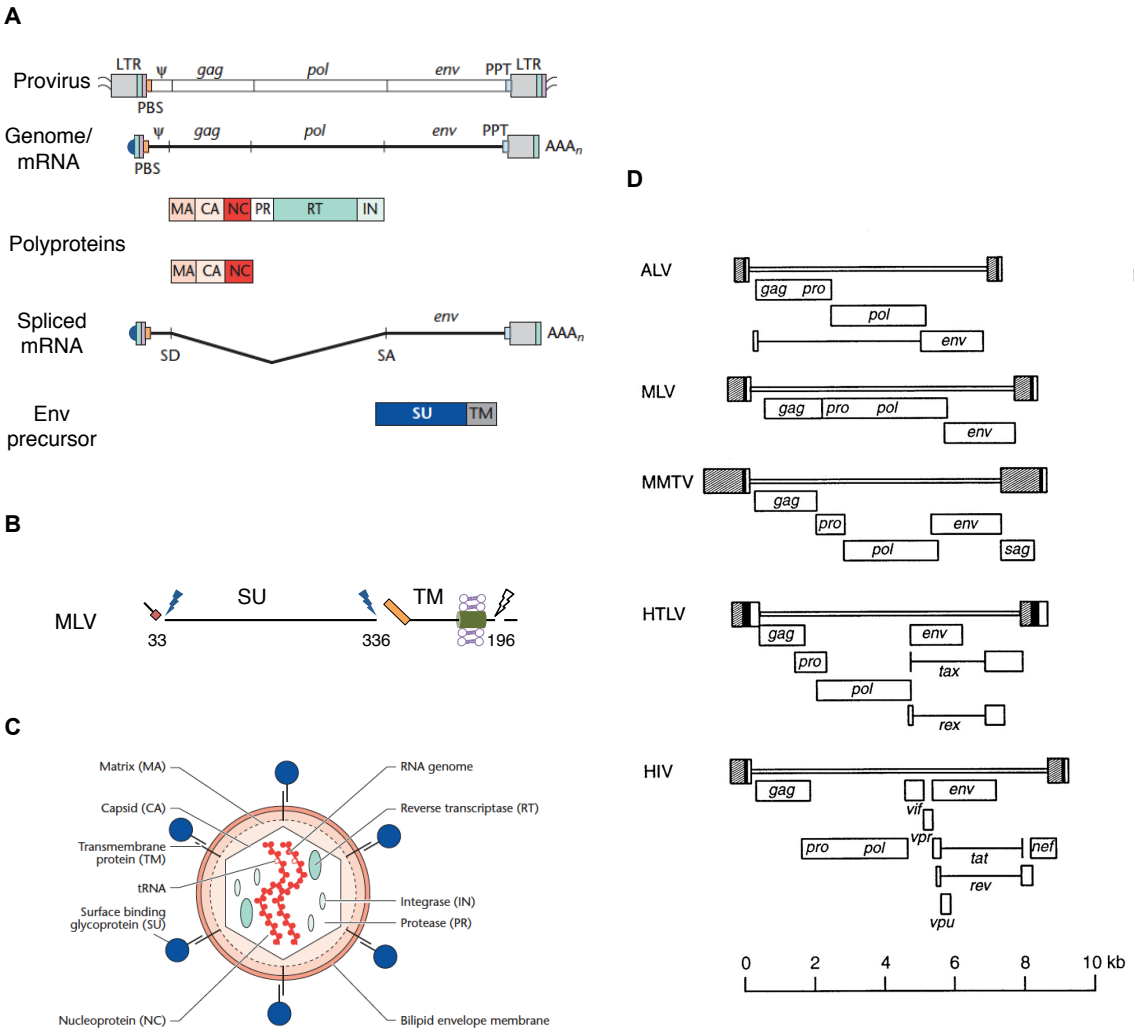


Figure 1.1: Retroviral genome and particle structure.

Phylogenetic relationships, based on RT, reveal that the retroviral diversity cluster into three distinct clades, also termed classes (although this term does not refer to a taxonomic designation) encompassing seven genera (Fields et al., 2001; International Committee on Taxonomy of Viruses. et al., 2012; Llorens et al., 2008) (Figure 1.2): (i) Epsilonretriviruses (class I, prototype: Walleye Dermal Sarcoma Virus (WSDV)) are complex retroviruses characterized by a C-type morphology (symmetrical spherical inner core) that infect fish and reptile species. (ii) Gammaretroviruses (class I, prototype: Murine Leukemia Virus (MLV)) constitutes the largest genus known, with members being simple retroviruses that infect mammals, birds and reptiles, and show a C-type morphology. (iii) Lentiviruses (class II, prototype: Human Immunodeficiency Virus (HIV)) are complex retroviruses characterized by having cylindrical or conical cores that infect mammalian species. (iv) Deltaretroviruses (class II, prototype: Human T-cell Lymphotropic Virus (HTLV) comprise a few complex retroviruses with a C-type morphology that infect some mammalian species. (v) Alpharetroviruses (class II, prototype: Avian Leukemia Virus (ALV)) are simple retroviruses with a C-type morphology that infect a wide range of avian species. (vi) Betaretroviruses (class II, prototype: Mouse Mammary Tumor Virus (MMTV)) are simple retroviruses characterized by a B-type (round eccentric core) or D-type (cylindrical core) morphology that assembles in the cytoplasm and have members infecting rodents, primates and sheep. (vii) Spuma- and spuma-like retroviruses (class III, prototype: Simian Foamy Virus (SFV) and Human

Endogenous Retrovirus-L (HERV-L) respectively) comprise a broad genus that contains the spumaviruses, which are complex retroviruses with a central but uncondensed core and, in contrast with the rest of the retroviruses, are able to undergo reverse transcription prior to infection (therefore having double stranded DNA (dsDNA) in the viral particle). Spuma-like retroviruses are simple endogenous retroviruses (ERVs) that are distantly related to spumaretrovirus and in some cases lack the *env* gene resulting in an entirely intracellular replication (ERV-L lineage) (Benit et al., 1999).

In order to enter a cell, all retroviruses require an interaction between the Env (SU) protein and a host cell surface molecule, the receptor (Figure 1.3A). This interaction is highly specific as minor changes in the receptor binding site can completely block viral infection (Albritton et al., 1993). The conservation, expression and distribution of suitable surface receptors determine the tissue and species tropism of the virus, as well as its pathogenesis. Retroviral receptors vary in their nature, structure and cellular function depending on the virus group. There are single-pass or multiple-pass transmembrane receptors (Albritton et al., 1989; Dalgleish et al., 1984), as well as secreted (Anderson et al., 2000) and glycosylphosphatidylinositol (GPI) anchored receptors (Rai et al., 2001) (Figure 1.4). Additionally, some retroviruses require more than one molecule (co-receptors) for entry (Overbaugh et al., 2001). Once it is correctly folded, glycosylated and oligomerized (in the ER), the single-pass transmembrane Env precursor is capable to interact with its receptor if encountered along the

secretory pathway. This interaction prevents the transport and expression of the receptor at the cell surface (Delwart and Panganiban, 1989; Matano et al., 1993; Nethe et al., 2005), and might lead to its degradation (Coffin et al., 1997). This phenomenon is known as receptor interference (or blockage) and plays a key role in superinfection resistance, where the production of an Env protein from an initial viral infection interact with its receptor (within the secretory pathway or at the cell surface) and prevents superinfection by a second virus that uses the same receptor (Coffin et al., 1997; Nethe et al., 2005). This property allowed virologists to assigned viruses to specific subgroups depending on their receptor usage (Sommerfelt and Weiss, 1990). For MLV the determinants for the receptor usage lie largely in two variable regions in its SU protein (Battini et al., 1998; Battini et al., 1992), which results in four groups depending on its receptor: ecotropic (MLV-E, infecting mouse and rat cells), amphotropic (MLV-A, infecting mouse, human and other cells), xenotropic (MLV-X, infecting non-murine cells) and dualtropic or polytropic (MLV-P, infecting murine and mink cells) (Weiss and Tailor, 1995). Similarly HIV is classified into X4 tropic, R5 tropic or dual tropic depending on its co-receptor usage (Berger et al., 1999). The determinants for different co-receptors are largely found in the V3 loop of the SU protein (Hwang et al., 1991; Shioda et al., 1991).

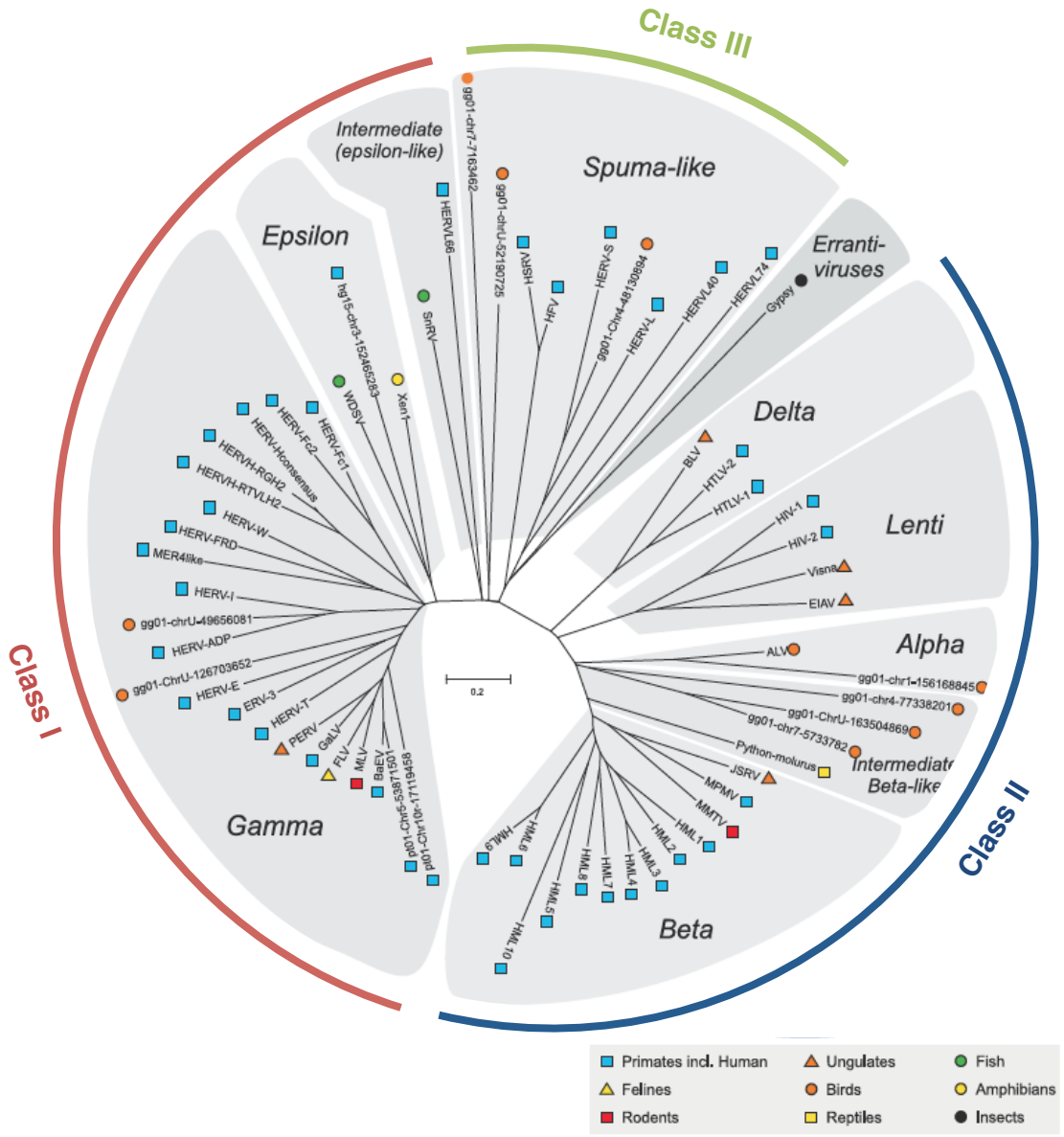


Figure 1.2: Retroviral diversity and classification.

Unrooted Pol neighbor joining (NJ) dendrogram of the seven retroviral genera and major classes (including endogenous and exogenous retroviruses). Phylogenetic location of the Gypsy transposon from the *metaviridae* family of retroviruses (errantiviruses) is also shown. Host species are indicated by symbols. Adapted from Jern et al., 2005 to include the classes described in Llorens et al., 2008.

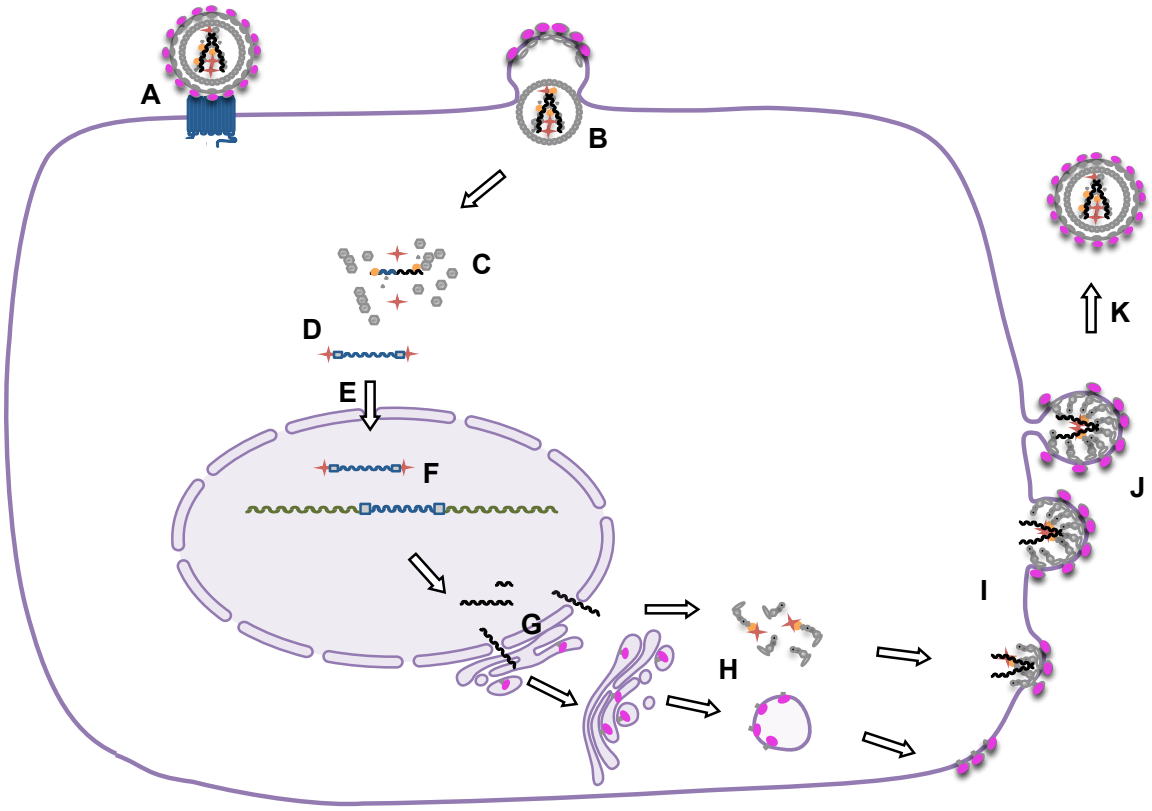


Figure 1.3: Steps in the retroviral life cycle.

Main steps in the retroviral life cycle from entry to viral production and maturation. **(A)** Receptor interaction, **(B)** membrane fusion and entry, **(C)** uncoating, **(D)** reverse transcription, **(E)** nuclear import, **(F)** integration, **(G)** mRNA/gRNA expression and export, **(H)** protein synthesis, **(I)** virion assembly, **(J)** budding and **(K)** maturation. Different steps are described in the main text. Gag and PR: gray forms. Pol: IN (red stars), RT (orange circles). Env: SU (Pink ellipse), TM (gray stem). A hypothetical multi-pass receptor is colored in blue. Viral RNA and DNA are colored in black and dark blue respectively. Adapted from an original figure kindly provided by Dr. Theodora Hatziioannou.

In contrast to other retroviral genera, all gammaretroviruses receptors have been identified to be multi-transmembrane solute transporters (Tailor et al., 2003) (Figure 1.4). This is the case of CAT-1 (cationic aminoacid transporter 1)(Albritton et al., 1989), a 14 transmembrane glycoprotein that mediates the transport of basic amino acids through the cell membrane and was identified as the MLV-E receptor (Figure 1.4A). Functional (amino acid transporters) CAT-1 orthologs are expressed at the cell surface of diverse animal species but specific residues at the third extracellular loop (envelope binding site) and particular glycosylation sites are responsible for the species tropism (Albritton et al., 1993; Wang et al., 1996). XPR1, an 8 transmembrane glycoprotein that mediates phosphate export from the cell (Legati et al., 2015), was found to be the receptor for both MLV-X and MLV-P (Battini et al., 1999; Tailor et al., 1999; Yang et al., 1999) (Figure 1.4A) and, as for CAT-1, a couple extracellular residues mediate recognition or resistance to MLV-X and/or MLV-P (Marin et al., 1999). ASCT2 (also known as SLC1A5 or RDR) is an exceptional 10 transmembrane sodium-dependent amino acids transporter that has been identified as the receptor for highly divergent gammaretroviruses (including some ERVs) and some betaretroviruses (Blond et al., 2000; Overbaugh et al., 2001; Sommerfelt and Weiss, 1990) (Figure 1.4A). Similarly, the Pit1 (a 12 transmembrane phosphate transporter) receptor is shared between GaLV (gibbon ape leukemia virus), WMSV (woolly monkey sarcoma virus) and some subtypes of FeLV (feline leukemia virus) (Overbaugh et al., 2001), whereas the related Pit2 glycoprotein

was determined as the receptor for MLV-A (Miller et al., 1994; van Zeijl et al., 1994) (Figure 1.4A). Other transporter multi-transmembrane gammaretrovirus receptors have been identified including for some ERVs (Soll et al., 2010; Tailor et al., 2003) (Figure 1.4A).

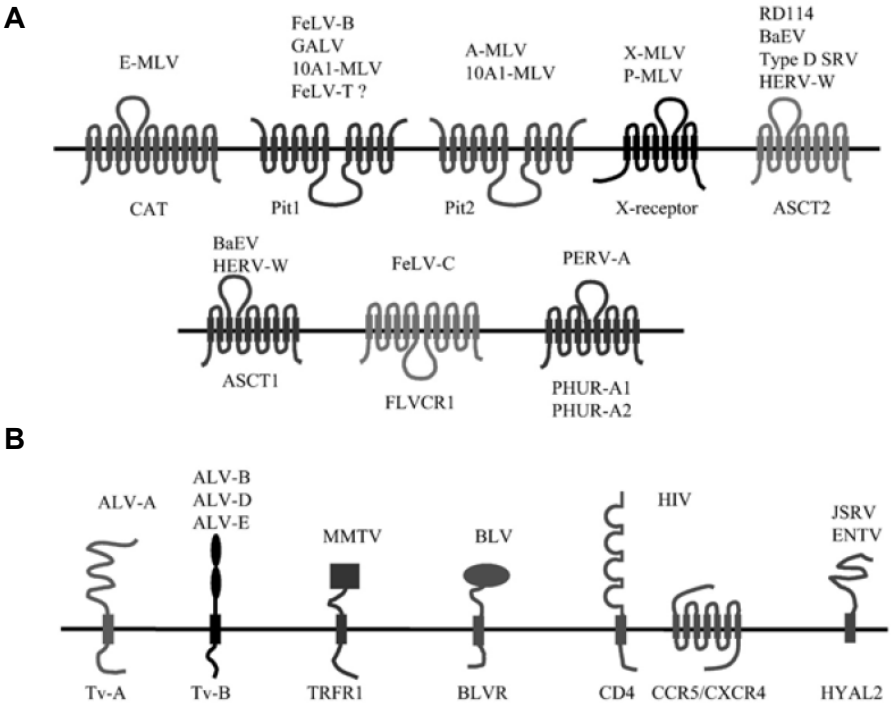


Figure 1.4: Retroviral Receptors

(A) Gammaretrovirus receptors. **(B)** Other retrovirus receptors. In all cases the protein name and the virus name are indicated below and above of each receptor diagram. For HIV both the receptor (CD4) and co-receptor (CCR5/CXCR4) are shown. Image adapted from Tailor et al., 2003.

HIV and other related lentiviruses typically require two receptor molecules for viral entry: the single transmembrane receptor CD4, that mediates binding to the SU protein (Dalgleish et al., 1984; Klatzmann et al., 1984; Sattentau et al., 1988), and a multi-transmembrane co-receptor CCR5 (R5 tropic) or CXCR4 (X4 tropic) (Alkhatib et al., 1996; Feng et al., 1996) (Figure 1.4B). Binding to CD4 leads to a conformational change that exposes the binding site to the co-receptor, which functions as a fusion receptor (Salzwedel et al., 2000). CCR5 and CXCR4 are chemokine receptors, while CD4 plays an essential role in the adaptive immune response (DeFranco et al., 2007) and is primarily expressed on helper T-cells, macrophages, and dendritic cells (Maddon et al., 1986) therefore explaining the limited cell tropism of HIV. Intriguingly and correlated with disease progression, it has been documented that HIV can shift its co-receptor usage resulting in differential $CD4^+$ T cell depletion in specific cellular compartments (Connor et al., 1997; Ho et al., 2005).

Although a pH-dependent endocytic pathway seems to be required for some betaretroviruses (Wang et al., 2008), alphaviruses (Padilla-Parra et al., 2014) and spumaviruses (Picard-Maureau et al., 2003), membrane fusion for most retroviruses occurs at the plasma membrane without the need of an acidic environment. They undergo complex conformational changes upon receptor binding resulting in the SU–TM complex dissociation that expose the fusion peptide, leading to the invasion of the target membrane and membrane fusion (Coffin et al., 1997; Pedersen et al., 2011) (Figure 1.3B). The events following

membrane fusion are still not completely clear, however in order to continue with infection, the retroviral cores need to disassemble in a process known as uncoating (Figure 1.3C). For lentiviruses some studies suggest that the core does not disassemble until it docks at the nuclear pore (Arhel et al., 2007), while others suggest that the uncoating of retroviral cores occurs along the transport of the viral genome to the nucleus and is stimulated by reverse transcription (Hulme et al., 2011; McDonald et al., 2002). Furthermore HIV CA (thus some kind of a core) seems to play a role in nuclear import and infection in non-dividing cells by its interaction with various nuclear transport proteins and the nuclear pore (Matreyek and Engelman, 2013). Whichever the model there is not a clear distinction between the timing of uncoating, reverse transcription and nuclear import (Figures 1.3C, D and E) (Campbell and Hope, 2015). In contrast to lentiviruses, gammaretroviruses are cell-cycle dependent and require the dissociation of the nuclear membrane during mitosis in order to access the nuclear DNA (Roe et al., 1993).

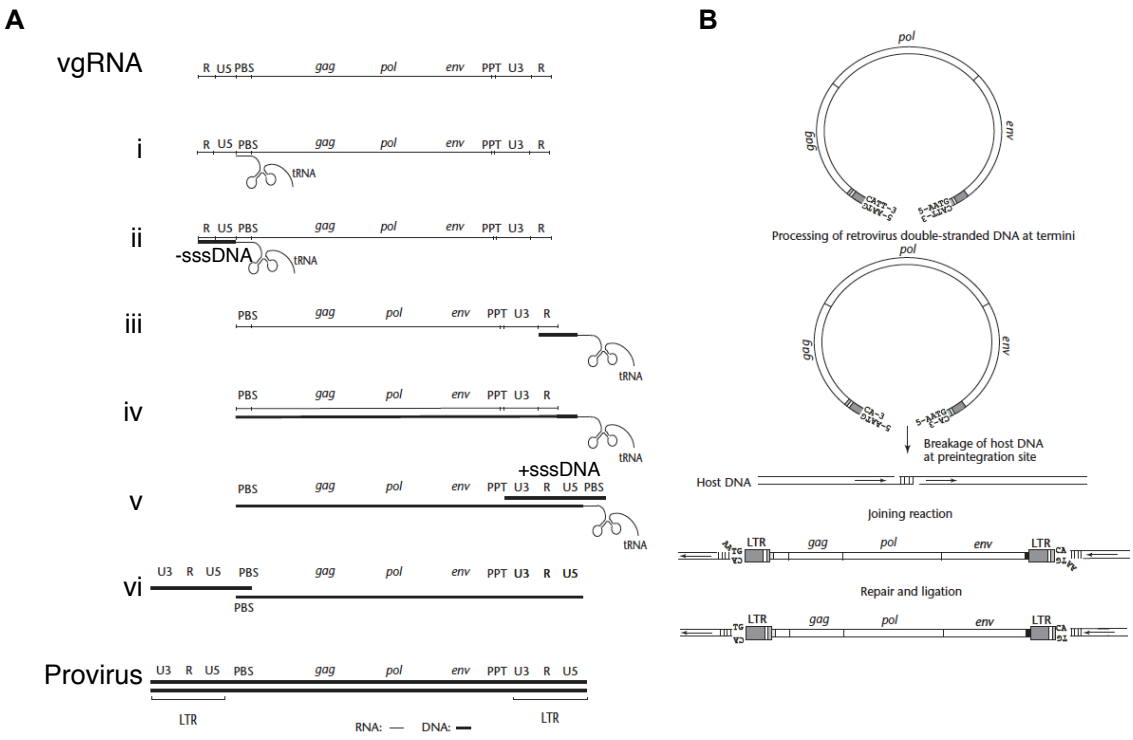


Figure 1.5: Retroviral reverse transcription and Integration.

(A) Reverse transcription process. Steps for the RT-dependent synthesis of the dsDNA provirus from a ssRNA viral genome (vgRNA). Diverse viral genomic features are indicated as well as the tRNA primer. Steps (i-vi) are described in the main text. RNA: light black line. DNA: thick black line. **(B)** Integration process. IN-mediated cleavage and ligation of dsDNA produced in (A) and host DNA. Subsequent repair by host enzymes results in a fully integrated provirus. The short direct repeat flanking the provirus is shown by vertical lines (IIII). Images adapted from Pedersen et al., 2006.

As infection proceeds, the viral RNA genome needs to be reverse transcribed into dsDNA by the action of the viral RT (Baltimore, 1970; Temin and Mizutani, 1970)(Figures 1.3D and 1.5A). Briefly: (i) A partially unfolded tRNA (different tRNA primers are used by distinct retroviruses (Lund et al., 1993)) anneals to the PBS of the viral genomic RNA (vgRNA) and (ii) leads to the synthesis of minus-strand DNA until the 5' end of the RNA genome, an intermediate termed minus-strand strong-stop DNA (-sssDNA). (iii) RNase-H degrades the RNA strand of the -sssDNA-vgRNA duplex leading to the first strand transfer, where the -sssDNA anneals to the R sequence on the cognate 3' end of the vgRNA (facilitated by NC (Rein, 2010)) and (iv) minus-strand DNA synthesis resumes. (v) RNase-H degrades the remaining RNA strand except for the duplex (RNA-DNA) corresponding to the PPT, which is resistant to RNase-H. The RNA portion of the PPT duplex serves as a primer for the plus-strand DNA synthesis that continues until a portion of the tRNA primer is reverse-transcribed, resulting in a DNA intermediate termed plus-strand strong-stop DNA (+sssDNA). (vi) RNase-H degrades the primer tRNA, leading to the second strand transfer, where the +sssDNA anneals to complementary sequences to the PBS at the 3' end of the minus-strand DNA (facilitated by NC). Subsequent plus and minus strand DNA synthesis results in a complete provirus structure with two LTRs (Coffin et al., 1997; Fields et al., 2001; Pedersen et al., 2011). Besides strand transfer to the ends of the viral genome, RT also shows frequent template shifting at internal positions when two heterologous sequences are packaged in the same viral

particle (DeStefano et al., 1992; Katz and Skalka, 1990). This internal template switching depends on the structure of the donor and acceptor templates and its favored by the stalling of the RT (Buiser et al., 1993; Duch et al., 2004). Co-infection by two or more viruses (or expression of ERVs) might lead to recombination through this internal template switching. In fact it has been observed that in multiple occasions, a gammaretroviral *env* gene recombined with other class I retroviruses, which might have facilitated cross-species transmissions (even among different vertebrate classes) by the change in their tropism (Henzy and Johnson, 2013). This ability to form "mosaic" species together with the RT error rate (for HIV-1 is 2×10^{-5} substitutions per site per replication cycle and it is similar for other retroviruses (Hu and Hughes, 2012)) have resulted in the high diversity observed for retroviruses.

After reverse transcription the newly synthesized viral dsDNA forms a complex with various viral and host proteins named the pre-integration complex (PIC) (Craigie and Bushman, 2014). Once inside the cell nucleus (discussed above), the PIC will guide the viral dsDNA to its integration into the host genome as an essential step in the viral life cycle (Figure 1.3F). The integration site preference seems to vary between retroviruses. Whereas lentiviruses preferred to integrate into active transcription units (Schroder et al., 2002), gammaretroviruses favor integration close to transcription start sites (Wu et al., 2003), while for other viruses such as alpharetroviruses integration seems to be almost random (Barr et al., 2005). The principal viral protein in the PIC is IN, which in addition to

mediating the enzymatic process of DNA integration, seems to be also a major determinant for integration site preference (Lewinski et al., 2006). Additional cellular co-factors, specifically LEDGF/p75 and BET proteins, have been implicated in the integration site preference by interacting with HIV and MLV IN, respectively, and tethering the viral intasome (IN in complex with viral dsDNA) to particular chromatin regions (Debyser et al., 2015). Once the intasome is docked to the targeted chromatin, IN catalyze the initial steps in the integration reaction (Figure 1.5B). Briefly, the dsDNA is first cleaved at the 3' end of each strand to form a two-nucleotide 5' overhang. Mediated by IN, the exposed 3' hydroxyl groups on the viral DNA promote a nucleophilic attack on phosphodiester bonds located in both strands of the target DNA and staggered by 4-6 bp (strand transfer reaction), resulting in the cleavage of host DNA and the joining of viral DNA 3' ends to the host genome (Kvaratskhelia et al., 2014). This reaction leaves DNA gaps at each host-proviral DNA junction that are repaired by cellular enzymes resulting in a 4-6 bp duplication of host DNA (Craigie and Bushman, 2014). Once integrated, the proviral DNA is then replicated together with cellular DNA during cell division and now evolves at the host substitution rate (assuming no selective pressures) (Feschotte and Gilbert, 2012). The inherent mutagenic potential of retroviral integration might result in the truncation or inactivation of host genes, as well as in the expression of tightly regulated genes, partially explaining the oncogenic capability of some retroviruses.

From its location in the host genome, the provirus is expressed and transcribed as a cellular gene. As described previously promoter and enhancer elements present in the U3 region of the 5' LTR initiate transcription of the vgRNA/mRNAs by the host RNAP II (Figure 1.3G). The RNAs are 5' capped (7-methyl-guanosine) and a poly(A) tail is added after particular sequences in the 3' LTR determining the end of the R region. In addition, complex retrovirus might encode transcriptional transactivators, such as HIV Tat, that bind to RNA structures in the 5' end of the RNA and stimulate RNAP II processivity, resulting in increased elongation of the transcripts towards the end of the provirus (Sodroski et al., 1985a; Sodroski et al., 1985b). Retroviral RNAs also contain splice donor and acceptor sites that are recognized by the cellular splicing machinery to produce the *env* and other accessory genes mRNAs. These spliced, partially spliced and complete mRNAs/vgRNAs need to be exported outside the nucleolus to start protein synthesis (Figure 1.3G). The cellular nuclear export pathway is highly selective letting only fully spliced mRNAs to reach the cytoplasm (Legrain and Rosbash, 1989). This phenomenon affects complex retroviruses that, in comparison to simple retroviruses, produce a range of unspliced, incompletely spliced and multiply spliced mRNAs (Figure 1.1D). For this reason some accessory genes of complex retroviruses also encode for proteins (such as HIV-1 Rev, HTLV-1 Rex, MMTV Rem and HERV-K Rec) that facilitate the nuclear export of unspliced or partially spliced viral RNAs and regulate their temporal expression (Mertz et al., 2009).

Whereas Env will be expressed, glycosylated, folded, oligomerized and processed as previously described, Gag and Gag-Pol polyproteins (as well as other accessory proteins) are translated in the cytoplasm by free ribosomes (Figure 1.3H). Expression of the Gag-Pol polyprotein requires the translational suppression of the Gag stop codon. Depending on the reading frame where Pol is encoded (Figure 1.1D), different retroviruses resolve this issue by read-through suppression (where a specific aminoacyl-tRNA is placed at the stop codon in the ribosome instead of the release factor, allowing for translation to resume) or frameshift suppression (where the ribosome undergoes a frame shift at a slippage region upstream of the stop codon, allowing for translation to resume in the alternative frame), mediated by specific RNA elements around the Gag stop codon (Pedersen et al., 2011).

The majority of retroviruses undergo particle assembly directly at the cell membrane (except for retroviruses with a B- and D- type morphology) (Fields et al., 2001) (Figure 1.3I). The Gag polyprotein pays the major role in assembly and is able to form particles even in the absence of *pol* or *env* (Campbell and Vogt, 1997; Delchambre et al., 1989). Additionally, Gag mediates the packaging of most of the particle components, including Pol (in the form of Gag-Pol polyprotein), vgRNA and other RNAs (Coffin et al., 1997; Kutluay et al., 2014). MA targets particle assembly to the plasma membrane through its N-terminal myristate group and a basic amino acid motif, which mediate a direct interaction with the plasma membrane and the binding to specific phospholipids (Saad et al.,

2006). Further Gag multimerization and possible relocation into lipid rafts and other membrane microdomains results in particle formation (Ono, 2010). As for the Env protein, which reached the plasma membrane though the secretory pathway, it is not completely clear how they get incorporated into the nascent particles. HIV-1 MA seems to interact and recruit Env proteins through its cytoplasmic tail, however this interaction is not essential for Env incorporation (Sundquist and Krausslich, 2012). Moreover, the process of Env incorporation is promiscuous and retroviruses can form (pseudotyped) viral particles carrying Env proteins from distantly related retrovirus or even from other viral families, arguing against specific determinants for Env incorporation. Another possibility that has been proposed is that both Gag and Env proteins are directed to the same microdomains in the plasma membrane (Briggs et al., 2003), however there is not a clear understanding of this process.

At this point the nascent viral particles need to detach from the cell membrane (Figure 1.3J). The budding process requires the recruitment of the host ESCRT machinery that stop Gag multimerization and catalyzes the excision of the cell and viral membranes (Sundquist and Krausslich, 2012). This recruitment is dependent on “late” domains encoded by retroviral proteins (and proteins form other viruses) that bind directly or indirectly to members of the ESCRT pathway (Bieniasz, 2006), leading to particle release. The remaining step in the retroviral life cycle is the maturation of the virions (Figure 1.3K), where the viral PR is liberated by autoproteolysis at a late stage of the viral assembly and mediates

the cleavage of the Gag and Gag-Pol precursors into their individual proteins (Pettit et al., 1998). This process leads to the assembly of CA to form the core structures (Ganser-Pornillos et al., 2008). In gammaretrovirus, PR also catalyzes the cleavage of the R-peptide at the Env C-terminal tail (Figure 1.1B), which activates the fusogenic potential of the TM protein (Rein et al., 1994). This maturation step perhaps represents a regulatory process by which retroviruses prevent reinfection of the producer cell (Pedersen et al., 2011). Indeed after this process the result is a fully infectious retroviral particle that has the ability to interact with its receptor on another cell and restart its life cycle (Figure 1.3).

This complex, and still not too entirely clear, series of events provide the host with several opportunities to counteract the detrimental effects of viral infection. In this regard the genetic conflict between viruses and their hosts has lead to the evolution of clever and efficient defense mechanisms by which both entities are trying to win the battle.

Retroviral Restriction Factors.

Of all the different antiviral mechanisms shown by the host, the first molecular lines of defense against viral infections are the restriction factors, which directly inhibit and disrupt essential steps for viral replication. In general, restriction factors share some common features (Blanco-Melo et al., 2012; Malim and Bieniasz, 2012): (i) They are dominantly and autonomously acting proteins that exhibit antiviral activity in simple cell-culture based assays. (ii) They are typically

expressed constitutively in some cell types, and/or are further induced by interferons (IFNs). (iii) They employ unique and unanticipated mechanisms to inhibit specific processes in viral replication. (iv) They have unusually diverse amino acid sequences as a consequence of antagonistic co-evolution with viruses (signatures of positive selection). (v) They can be antagonized by viral proteins or specific mutations.

The number of proposed restriction factors has exploded since the term was established in the early 1970s, when the expression of Fv1 was shown to protect against MLV infection (Lilly, 1970), and particularly since the discovery of the first restriction factor against HIV (Sheehy et al., 2002). Currently there are more than 20 proteins that have been proposed to serve as restriction factors against a variety of animal viruses (Kluge et al., 2015). For the purposes of this chapter a number of restriction factors were selected that have been of interest to subsequent studies in this thesis (Table 1.1).

Table 1.1: Selected Retroviral Restriction Factors

Restriction Factor	Mechanism of restriction	Life cycle step disrupted (*)	Viral antagonism	Induced by IFN	Positive selection
APOBEC3	Hypermutation	Reverse Transcription (D)	Vif (lentiviruses), Bet (spumaviruses), GlycoGag (MLV)	Yes	Yes
TRIM5	Viral Core Disruption and signaling	Uncoating and reverse transcription (C,D)	CA	Yes	Yes
SAMHD1	Depletion of dNTP pools and degradation of vgRNA	Reverse Transcription and RNA expression (D,G)	Vpx (HIV-2, SIV), Vpr (SIV)	Yes	Yes
Tetherin	Tethers nascent virions	Budding (J)	Vpu (HIV-1, SIV), Nef (SIV, HIV-1), Env (HIV-2, SIV)	Yes	Yes
Mx2	Blocks nuclear import and affects stability of nuclear viral DNA	Nuclear Import and Integration (E,F)	CA (HIV, SIV)	Yes	Yes
SERINC5,3	Reduces membrane fusion	Entry (B)	Nef (HIV, SIV), GlycoGag (MLV), S2 (EIAV)	No	No
MOV10	RNA processing, inhibition of reverse transcription	Reverse transcription and Protein synthesis (D,H)	NI	Yes	No
Fv1	Viral Core Disruption	Uncoating and integration (C,E)	CA (MLV)	No	Yes
Fv4	Receptor interference	Receptor interaction (A)	NI	No	NI
enJSRVs	Receptor interference and inhibition of particle release	Receptor interaction and budding (A,J)	NI	No	NI

* Letters in parenthesis refer to specific steps in the retroviral life cycle depicted in Figure 1.3
NI: Non Identified
Table adapted from Kluge et al., 2015

(i) APOBEC3: APOlipoprotein B Editing Catalytic subunit-like 3 (APOBEC3) G was first characterized as the restriction factor antagonized by the HIV-1 Vif accessory protein (Sheehy et al., 2002). It is incorporated into virions and exhibits antiviral activity in target cells by catalyzing the deamination of 2'-deoxycytidine (dC) to 2'-deoxyuridine (dU) on minus-strand DNA during reverse transcription, mediated by its cytidine deaminase domain (Yu et al., 2004). This change leads to guanine (dG) to adenosine (dA) mutations in the proviral DNA (APOBEC3-mediated hypermutation), rendering it replication defective. The human genome encodes for seven APOBEC3 proteins from which APOBEC3A, APOBEC3B, APOBEC3D, APOBEC3F, APOCE3G and APOBEC3H have been shown to significantly decrease primate lentivirus infection (Bishop et al., 2004), through hypermutation and deamination-independent mechanisms (Gillick et al., 2013; Holmes et al., 2007). HIV-1 Vif binds to several APOBEC3 proteins and targets them for ubiquitinylation and degradation resulting in the down regulation of its antiviral activity (Goila-Gaur and Strebel, 2008). The mouse genome encodes only a single APOBEC3 protein and in addition to its hypermutation function, there are indications for a deamination-independent antiviral function targeting reverse transcription and antagonized by MLV glycoGag (Stavrou et al., 2013).

(ii) TRIM5: TRIM5α is a member of the tripartite motif (TRIM) family of proteins characterized by a RING domain (E3 ubiquitin ligase), one or two B-box domains (required for higher-order assembly) and a coiled-coil domain (for

dimerization) (Blanco-Melo et al., 2012; Zheng et al., 2012). TRIM5α recognizes CA in incoming retroviral cores and assembles into a three-dimensional lattice around them (Ganser-Pornillos et al., 2011). The variable C-terminal B30.2/SPRY domain in TRIM5α is the determinant for the recognition and binding to particular retroviral CA, and this interaction might lead to the premature disassembly of viral cores and the degradation of its components before reverse transcription is complete (Stremlau et al., 2004; Stremlau et al., 2006). Additionally, TRIM5α has been shown to promote innate immune signaling upon interaction with the cores (Pertel et al., 2011). Surprisingly, in owl monkey and macaques, independent LINE-mediated retrotransposition events led to the replacement of the TRIM5 SPRY domain for a cDNA expressing cyclophilin A (CypA) (Nisole et al., 2004; Sayah et al., 2004; Virgen et al., 2008; Wilson et al., 2008). This TRIMCyp fusion protein exploits the conserved interaction of lentiviral CA to CypA (Goldstone et al., 2010), resulting in a potent lentiviral inhibitor. Several CA mutants have been identified that antagonize the effects of TRIM5α and TRIMCyp (Chatterji et al., 2005; Veillette et al., 2013).

(iii) SAMHD1: SAM domain and HD domain-containing protein 1 (SAMHD1) hydrolyzes deoxynucleoside triphosphates (dNTPs) to deoxynucleosides leading to the reduction of the intracellular dNTPs concentrations and the inhibition of reverse transcription (Lahouassa et al., 2012). Mutations in SAMHD1, as well as other cellular nucleases, were initially associated with

Aicardi-Goutières syndrome (AGS), a hereditary autoimmune disease that is characterized by aberrant up-regulation of type I IFN responses (Rice et al., 2009). Recently, SAMHD1 has also been shown to have a nuclease activity that degrades different retroviral vgRNAs (Choi et al., 2015; Ryoo et al., 2014). Thus SAMHD1 might also function to regulate the accumulation of intracellular endogenous nucleic acids (perhaps expressed from ERVs and retrotransposons), which in AGS has been proposed to trigger a type I IFN response (Stetson et al., 2008). Lentiviral Vpx proteins (and some Vpr) have the ability to induce degradation of SAMHD1 by a similar mechanism to the degradation of APOBEC3G by Vif (Hrecka et al., 2011; Laguette et al., 2011; Lim et al., 2012).

(iv) Tetherin: also known as BST-2, CD317, or HM1.24 is a restriction factor that traps nascent virions at the cell surface from a broad spectrum of enveloped viruses (Neil et al., 2008; Van Damme et al., 2008). Tetherin is a membrane glycoprotein comprised of a short N-terminal cytoplasmic tail, a single pass transmembrane helix (TM), a helical coiled-coil ectodomain (CC) that drives parallel homodimer formation, and a C-terminal glycosylphosphatidylinositol membrane anchor (GPI). This highly unusual architecture, rather than primary sequence, is critical for Tetherin function (Perez-Caballero et al., 2009), thus Tetherin likely traps virions simply by the partitioning of linked membrane anchors between virion and cell membranes, from where they may be endocytosed. Interestingly, axially configured Tetherin dimers trap virions

primarily through the insertion of their GPI anchors into the lipid envelopes of budding virions (Venkatesh and Bieniasz, 2013). Remarkably, three different primate lentiviral proteins have acquired the ability to antagonize Tetherin: Vpu (HIV-1 and some SIVs), Nef (SIV) and Env (HIV-2) by recognizing distinct Tetherin epitopes (Jia et al., 2009; Le Tortorec and Neil, 2009; McNatt et al., 2013; Zhang et al., 2009). Interestingly, pathogenic revertants of the normally non-pathogenic SIVmac (Δ*nef*) arise in rhesus macaques, in which SIVmac Env has also adapted to antagonize Tetherin (Serra-Moreno et al., 2011). In addition to its minimalistic antiviral mechanism, Tetherin has also been implicated to promote innate immune signaling by detecting HIV-1 particles (Galao et al., 2012).

(v) Mx2: also known as MxB, has been proposed to prevent the nuclear import and integration of proviral DNA, through a mechanism not yet completely understood (Goujon et al., 2013; Kane et al., 2013; Liu et al., 2013). Human Mx2 primordially inhibits lentiviruses and co-localizes with nuclear pore components at the nuclear envelope (Goujon et al., 2013; Kane et al., 2013). It appears that as for TRIM5α, CA is the determinant for Mx2 activity and CA mutants have been identified that antagonize the effects of Mx2 (Busnadiego et al., 2014).

(vi) SERINC5 & 3: Although the antiviral mechanism is still elusive, SERINC5 and SERINC3 are incorporated into virions and block the complete delivery of their viral cores into the cytoplasm during viral entry (Rosa et al., 2015; Usami

et al., 2015). Nef (SIV and HIV) and MLV glycoGag antagonize the effects of SERINC5 and SERINC3 by preventing their incorporation into virions (Rosa et al., 2015; Usami et al., 2015).

(vii) MOV10: MOV10 activity has been contradictory and a clear mechanism of action is still pending. MOV10 has been implicated in maintaining genome integrity of germ line cells by silencing ERVs and transposon transcripts (Frost et al., 2010). As such, MOV10 was found to be the mammalian ortholog of plant and insect RNA silencing effector proteins (Zheng et al., 2012). In somatic cells, MOV10 is also able to restrict retrotransposition of endogenous retroelements, but has no effect on exogenous retroviruses (Arjan-Odedra et al., 2012). However, and with contradictory results, other studies have shown that MOV10 is also packaged into virions (through interaction with Gag and RNA) where it has been proposed to interfere with viral reverse transcription in target cells (Abudu et al., 2012; Burdick et al., 2010). Additionally, MOV10 also localizes to mRNA processing bodies (involved in the storage and decay of some RNA species) suggesting a role in mRNA decay (Gallois-Montbrun et al., 2007).

(viii) Fv1: like TRIM5α, the precise mechanism of Fv1 restriction is unclear but this protein also has the ability to recognize the CA protein of incoming retroviral cores (particularly MLV) and interfere with viral integration after reverse transcription (Hatziioannou et al., 2004; Hilditch et al., 2011; Jolicoeur and Rassart, 1980; Yang et al., 1980). Remarkably, Fv1 was the first

documented case that ERV sequences can be co-opted as antiviral genes, in this case from an ERV *gag* gene related to the Murine ERV-L (MuERV-L) (Best et al., 1996; Yan et al., 2009). Fv1 can also inhibit HIV-1 infection upon fusion to CypA impairing its integration (Schaller et al., 2007).

(ix) Fv4: similar to Fv1, the Fv4 protein was also co-opted from an ERV, this time from an *env* gene closely related to the ecotropic MLV, which potently blocks MLV infection (Inaguma et al., 1991; Odaka et al., 1980). Fv4 contains a single aminoacid substitution in its fusion peptide that destroyed its fusogenic potential, while still retaining the ability to interact with its receptor (mCAT1) (Taylor et al., 2001; Yamaguchi et al., 2003). These results suggest that Fv4 was co-opted as an antiviral protein by exploiting the receptor interference mechanism showed upon superinfection resistance by exogenous infections, as discussed above. It is estimated that the co-option of this protein occurred ~500,000 years ago (Kozak, 2015).

(x) enJSRVs: as with Fv4, sheep and goats have co-opted recent ERVs derived from the betaretrovirus Jaagsiekte sheep retrovirus (JSRV). There are at least 20 proviral loci closely related to JSRV (endogenous JSRVs or enJSRVs)(Arnaud et al., 2007), many of which have retained *gag* and *env* genes with full coding potential. Therefore, enJSRVs can inhibit exogenous JSRV infections by two distinct mechanisms, one by receptor interference, by expressing inactive *env* proteins (Spencer et al., 2003), and by expressing Gag proteins that act in a transdominant fashion. This is achieved by forming

particles together with exogenous JSRV Gag in the cytoplasm that are defective in particle release (Murcia et al., 2007). At least one enJSRV provirus expressing a complete *env* gene is estimated to be ~3 MY old, and is fixed in the domestic sheep population (Varela et al., 2009).

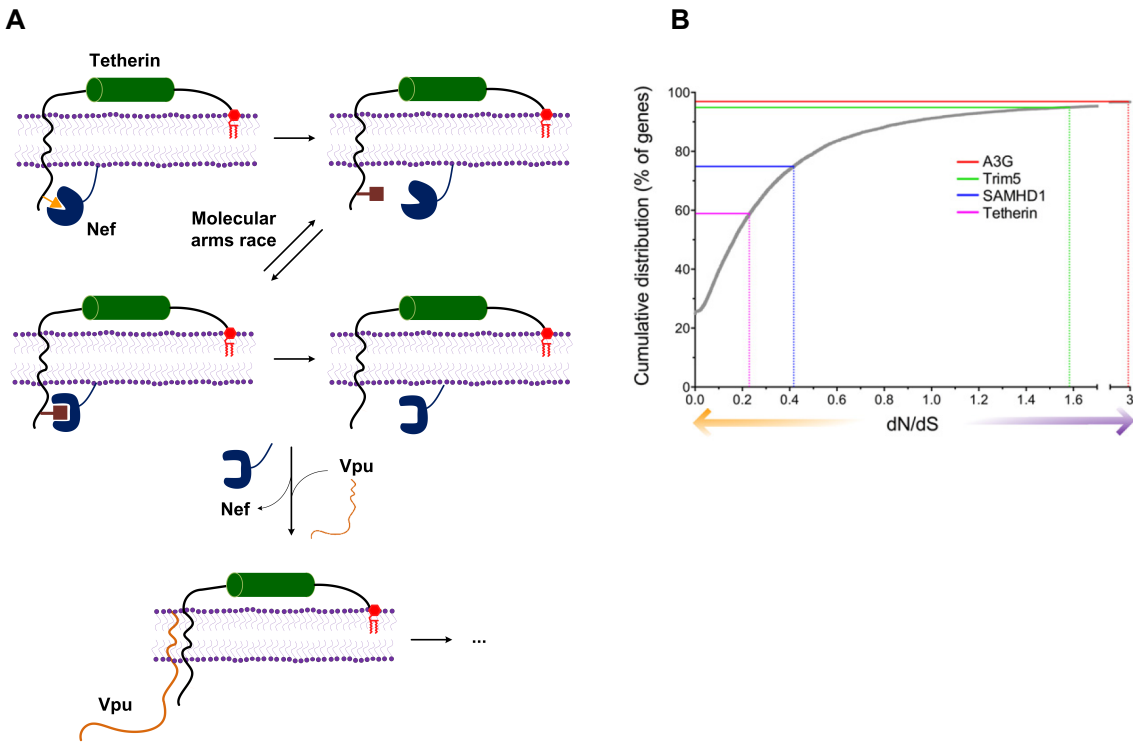


Figure 1.6: Molecular arms race between virus and host and its effect on host restriction factors.

(A) Molecular arms race between a restriction factor (Tetherin) and lentiviral accessory proteins. Nef proteins of SIVs antagonize Tetherin by interacting with the Tetherin cytoplasmic tail, the genetic conflict posed by this interaction embarks both proteins into a molecular arms race that, in this case, led to the evolution of Vpu, as primate lentiviruses were transmitted between species. Colored figures indicate Tetherin sequences in the cytoplasmic tail that are recognized by Nef and are hence rapidly evolving under positive selection. **(B)** Cumulative frequency distribution of dN/dS ratios for 12,404 Human-Chimpanzee orthologous gene pairs. Adapted from previously computed data (Consortium, 2005). Strict positive selection (dN/dS>1) and purifying selection (dN/dS<1) are indicated by purple and orange arrows respectively. The dN/dS value for each restriction factor is indicated by the dotted lines. The solid lines indicate the percentage of orthologous gene pairs with lower dN/dS ratios. Images adapted from Blanco-Melo et al., 2012.

It is clear that restriction factors target key steps of the viral life cycle that are essential, and in order for the virus to replicate in a particular host it had to acquire specialized countermeasures against these antiviral proteins. It is likely that this antagonistic coevolution represents a major driver for evolutionary change in both viruses and their hosts. Restriction factors variants that confer an advantage are selected by detrimental viral infections and could be rapidly fixed in the host population. On the other hand, the reduction in viral fitness in the newly adapted host requires the selection of viral variants that have acquired mutations or new functions that relieve restriction and restore fitness. Iterative cycles of this genetic conflict constitute a molecular “arms race” and can result in the rapid evolution of restriction factors and their viral targets or antagonists (Blanco-Melo et al., 2012; Daugherty and Malik, 2012) (Figure 1.6A).

A molecular arms race between protein-coding genes is often identified by an observed abundance of mutations that change the amino acid sequence (non-synonymous mutations) over those that do not (synonymous mutations). Protein sequences that show this imbalance are thought to have evolved under positive selection, where the sequence diversification is justified by the fixation of beneficial alleles. This type of selection contrasts with purifying selection (abundance of synonymous over non-synonymous mutations), where fixation of alleles is driven by the need to preserve protein function through the elimination of deleterious mutations. A simple numerical way to distinguish between both types of selection is by calculating the ratio of non-synonymous mutations over

all potential non-synonymous sites (dN) and compare it to the ratio of synonymous mutations over all potential synonymous sites (dS). In alignments of genes, portions of genes, or individual codons, a dN/dS > 1 would be indicative of positive selection, whereas a dN/dS <1 is indicative of purifying selection. As intuitively expected, the majority of human genes have evolved under purifying selection (dN/dS <1) with a small subset of genes that have evolved under positive selection (Figure 1.6B) (Blanco-Melo et al., 2012; Consortium, 2005; Meyerson and Sawyer, 2011). Genes that exhibit signatures of positive selection include those involved in sensory perception, likely driven by temporal or migration-induced changes in the need to sense the environment, food, or location of predators. Predictably, positively selected genes also include those involved in immune responses and pathogen defense (Kosiol et al., 2008), including restriction factors. The entire coding sequence of APOBEC3G and TRIM5α are among the highest dN/dS ratios of all human genes (Sawyer et al., 2004; Sawyer et al., 2005; Song et al., 2005) (Figure 1.6B). Positive selection typically acts on domains or codons that participate in the interaction between proteins and their targets. Although the entire coding sequence of Tetherin and SAMHD1 show dN/dS < 1 (when comparing it to the orthologous chimpanzee sequence, Figure 1.6B), sequences that are targeted by viral antagonists are shown to have clear signatures of positive selection (Gupta et al., 2009; Lim et al., 2012; McNatt et al., 2009) (Figure 1.6A), and similar results have been observed for other restriction factors (Table1.1).

Positive selection has caused high protein sequence variability in restriction factors from different species. Consequently, viral adaptation to antagonize or evade a particular restriction factor variant in one host species can come at the cost of susceptibility to variants of the same restriction factor in another potential host. Thus, antagonistic co-evolution of virus and a particular host can reduce the probability of an individual virus to evade (or antagonize) the defense mechanisms that will confront it when the opportunity to colonize a new host species appears. It should be straightforward at this point to notice that although restriction factors are most frequently studied because of their ability to inhibit modern, clinically important retroviruses, their existence (and its current sequence) in modern genomes is the result of molecular arms races and selective pressures imposed by viruses in the past.

While the aforementioned restriction factors share some common properties, their mechanisms of action and evolutionary origins are quite different from each other. How did such a diverse array of antiviral proteins arise? Or equally daunting, how do new genes and functions arise in general? In principle, new functions could be generated through minor adaptation of cellular genes whose products already have an intrinsic capacity to perform such a function. Alternatively, they might originate *de novo* as new genes with trace side activities or innovations that ultimately become the selected new function (Bergthorsson et al., 2007). Moreover, the fate of any gene depends on its ability to preserve its function (robustness) while accumulating mutations to adapt to a changing

environment (evolvability). Although initially contradictory, the interplay between this two forces (robustness and evolvability) might allow evolving genes to sample different variants while resisting potentially deleterious mutations (Masel and Trotter, 2010; Wagner, 2011). In either case, gene duplication provides the genetic “raw material” that might facilitate genotype sampling (by releasing evolutionary constraints in the paralog gene) leading to the acquisition of new functions (Ohno, 1970).

The APOBEC3 family is likely derived from duplicated copies of cytidine deaminases (AID and APOBEC1) that have specific roles in editing cellular DNA and RNA. Thus, in this case, a normal cellular function was simply redirected to hypermutate viral genomes. In a similar manner, Mx2 was most certainly evolved from a duplication of the also antiviral *mx1* gene. One could imagine that the enzymatic regulation of cellular dNTP levels by SAMHD1 might have served some important regulatory function that was subsequently exploited by cells to inhibit the replication of retroviruses (and perhaps other DNA viruses). TRIM5 likely represents an intermediate example, whereby genes with some pre-existing, but mechanistically unrelated, function were mutated to a form with antiretroviral activity. Consistent with this idea, most of the dozens of TRIM proteins do not possess intrinsic antiretroviral activity. However, there are some that can exhibit weak antiretroviral activity when overexpressed (Yap et al., 2004; Zhang et al., 2006). Moreover, a variety of TRIM genes, that otherwise lack antiretroviral function, can be endowed with anti-HIV-1 activity if their C-terminal

SPRY domains are replaced with cyclophilin (Yap et al., 2006; Zhang et al., 2006). These data suggest that the architecture of TRIM proteins, perhaps a propensity to assemble hexameric lattices (Ganser-Pornillos et al., 2011), lends itself to the acquisition of antiretroviral activity.

In the case of Tetherin, it is difficult (albeit not impossible) to imagine a precursor gene with a similar function. There are no known cellular proteins that have related sequence or function, and so far *tetherin* appears to be an orphan gene present only in mammals and some reptiles (Heusinger et al., 2015). Moreover, Tetherin is not expressed in the vast majority of cells unless they are treated with interferon, and mice that lack a *tetherin* gene have no obvious deficiencies (Liberatore and Bieniasz, 2011). Every sequenced genome harbors a significant fraction of "orphan" genes whose origins are obscure due to the absence of sequence or functional similarity to other genes (Khalturin et al., 2009; Palmieri et al., 2014; Tautz and Domazet-Loso, 2011), thus addressing the evolutionary history of Tetherin certainly will be a challenging task.

Another recurrent source for genetic innovation is represented by the ERV sequences that have provided an evolutionary advantage for the host to keep. In the particular case of ERV *env* genes mediating receptor interference, examples of this mechanism have been documented in chickens (Robinson et al., 1981), sheep (Spencer et al., 2003), mice (Gardner et al., 1991; Odaka et al., 1980; Wu et al., 2005) and cats (Ito et al., 2013), emphasizing its simple but profound effect. Co-opted ERV sequences such as Fv1, Fv4, enJSRVs and others have

arisen independently at different times and in different species to deal with diverse viral infections, highlighting the key role that ERVs have in shaping the evolution of their hosts.

Endogenous Retroviruses.

About 8 and 10 percent of the human and mouse genomes, respectively (Lander et al., 2001; Mouse Genome Sequencing et al., 2002), are comprised of sequences of retroviral origin. These ERV sequences originated from an initial retroviral infection that was able to infect and integrate in the host germ line cells. This integrated retroviral genome was then vertically inherited to the offspring as a host allele. During thousands to millions of years, some of these sequences acquired inactivating mutations and were fixed in ancestral populations by genetic drift, while others became fixed by providing an evolutionary advantage to the host (Figure 1.7A). Since their discovery, numerous ERVs (mostly simple retroviruses) have being identified in a variety of animal species indicating that germ line retroviral infection events have occurred multiple times during the course of evolution (Hayward et al., 2013). This ERV diversity observed in modern animal genomes corresponds to germline retroviral insertions that survived natural selection and reached fixation, which is a significantly reduced set of all past retroviral infections. Nevertheless, this ERV diversity still represents a vast molecular “fossil record” of past retroviral infections.

Once within the germline, the endogenized provirus can proliferate giving rise to multi-copy lineages (or families) of related ERV sequences in the host genome (Bannert and Kurth, 2006). Such lineages are primordially named by adding one or two letters before the abbreviation ERV describing the host species in which they were initially identified. Additionally, based on the particular tRNA used to prime reverse transcription, a one-letter amino acid symbol for that tRNA is added to end of the name. For example, the human HERV-K lineage uses tRNA-Lys as a primer while the mouse MuERV-L uses tRNA-Leu.

Figure 1.7: Origins of ERVs and their mechanisms of expansion.

(A) Process of endogenization of a retrovirus. An initial retroviral infection that expanded through a population by horizontal transmission is fortuitously able to infect the germline cells of the host. From there the virus could be inherited vertically from generation to generation. Some of these retroviral sequences will be fixed in the population as cellular genes. Image adapted from Dewannieux and Heidmann, 2013. **(B)** Mechanisms used by ERVs to expand in the germline. Reinfection: ERVs that have maintained coding potential for all its genes (specially *env*) can make retroviral particles and reinfect germline cells (and also other somatic cells). Retrotransposition: ERVs that have lost its *env* gene can still re-integrate into the same cell through entirely intracellular mechanisms. Complementation *in trans*: Partially or completely inactive ERVs may continue its expansion by using the replication machinery of other endogenous or exogenous retroelements (indicated in red). Image adapted from Bannert and Kurth, 2006.

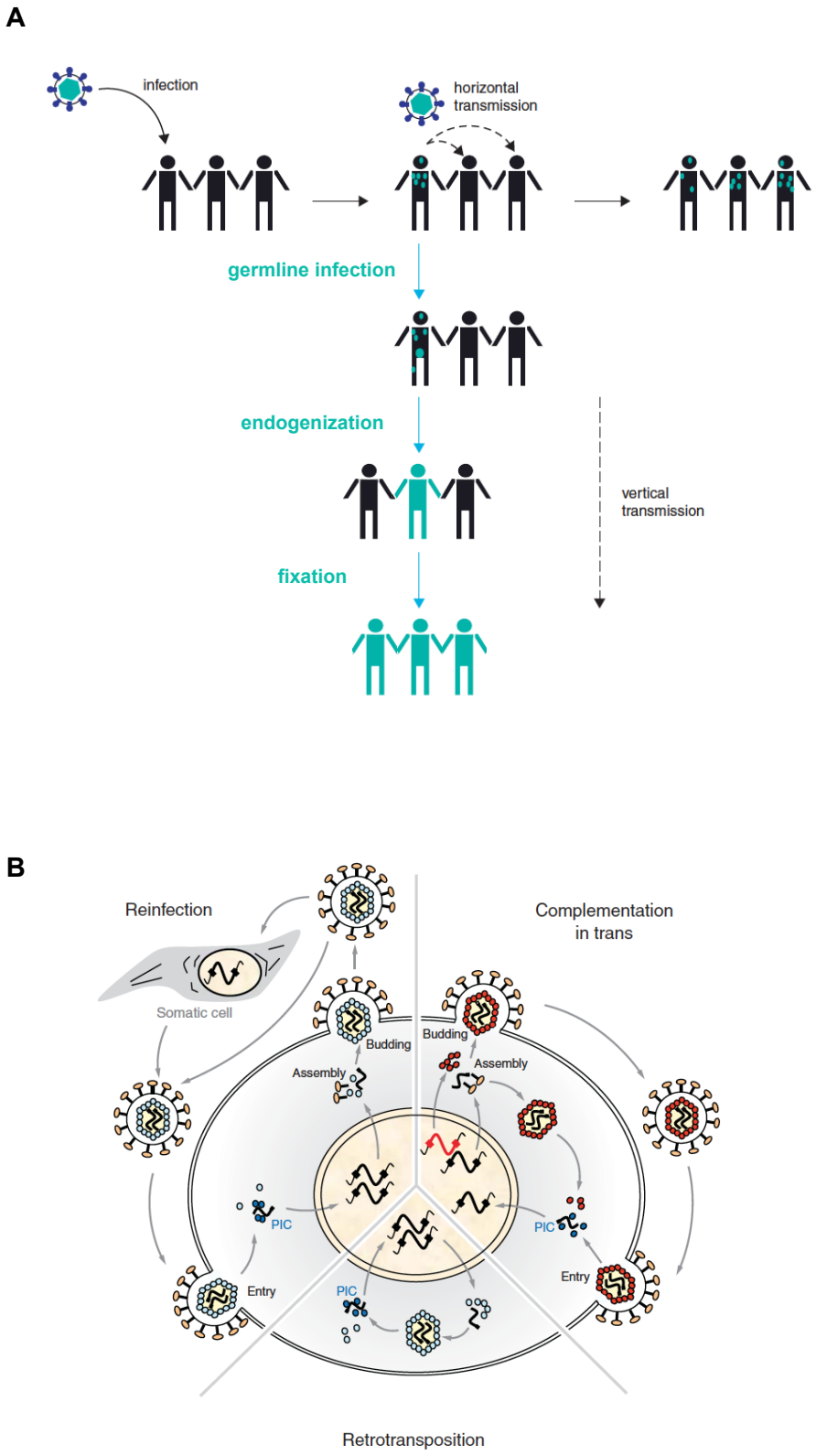


Figure 1.7: Origins of ERVs and their mechanisms of expansion.

The mechanism behind the expansion of a particular lineage is related to the proviral structure itself. If an ERV is still able to encode for all major proteins (Gag, Pol and Env), this sequence could expand in copy numbers by producing infectious particles that could (re-)infect other germ line cells (and possibly also some somatic cells) (Figure 1.7B, Reinfection). This is the case when a provirus has been recently acquired or is not yet completely endogenized (fixed in the population). Indeed, a recent study observed 36 HERV-K proviruses that are polymorphic among different human populations with two of them retaining full coding capacity for all of their genes (including the *rec* accessory gene), suggesting that they could still be infectious (Wildschutte et al., 2016). Another important example is that of the koala retrovirus (KoRV) that recently entered the germline (~100 years ago (Tarlinton et al., 2006)) and is present in variable copy numbers in north Australian koalas, but absent in some south Australian ones (Simmons et al., 2012). This mode of expansion is especially dependent on a functional *env* gene to secure (re-)entry into cells. It has been observed that some lineages with very large copy numbers lost its *env* gene and expanded by reintegrating in the same cell through completely intracellular mechanisms similar to retrotransposition (Figure 1.7B, Retrotransposition) (Magiorkinis et al., 2012). This amplification mechanism also requires the adaptation of Gag to assemble particles in the cytoplasm and not at the plasma membrane, as shown by studies that interchanged the mode of replication of related ERVs by swapping the N-terminal end of Gag between them (Ribet et al., 2008a). The success of this

mode of replication can be explained by an increased probability of reintegration if the “infectious” elements are present locally at high concentrations or they do not have to avoid targeting by the extracellular immune system (Dewannieux and Heidmann, 2013). Although there can be some advantages for the host to keep proviruses with coding potential (see below), the vast majority of the ERVs in animal genomes are filled with inactivating mutations, including deletions and insertions. These inactive proviruses can still proliferate by “borrowing” the required components from active retroviruses (endogenous or exogenous) or other retrotransposons (such as LINE elements), as long as they can still be transcribed and regulatory sequences are preserved (PBS, PPT, Ψ, etc.) (Figure 1.7B, Complementation *in trans*). This mechanism has also been widely used and in fact more than 60% of all HERV-W proviruses have been generated with the help of LINE elements (Pavlicek et al., 2002). Ultimately, the single most abundant ERV structure in animal genomes is just composed of one LTR (soloLTR) that arose by homologous recombination between the paired LTRs of complete proviruses (Belshaw et al., 2007) and it is in 10-fold excess compared to other ERV structures (Mager and Stoye, 2015). This internal deletion could further reduce the burden to the host genome and might have been preferentially fixed in the population by genetic drift (Dewannieux and Heidmann, 2013).

The potential for amplification of ERVs despite inactivation, impose an important pressure on the host to regulate their expression and minimize their effects. Although the role of ERVs in disease is still debatable, ERV transcription (and in

some cases protein expression) is often unregulated in certain cancers, autoimmune diseases and even in other retroviral infections (such as HIV) (Magiorkinis et al., 2013; van der Kuyl, 2012). As previously mentioned some restriction factors, such as APOBEC3G, SAMHD1, MOV10, Fv1, Fv4 or enJSRV, can act against ERVs and inhibit their detrimental effects, although their action seems to have been more critical to prevent the initial germline invasion (Johnson, 2015). Another way to keep ERVs under control is by tight epigenetic regulation of their expression through DNA methylation or histone modifications (Leung and Lorincz, 2012; Maksakova et al., 2008). In fact, epigenetic silencing is such an important mechanism for the control of distinct ERV lineages and retrotransponsons that a particular family of zinc finger proteins (KRAB-ZFB), which recruit histone methylation complexes and guide them to distinct ERVs, seems to be fast evolving and expanding in mammalian genomes, suggesting its involvement in a molecular arms race with past endogenous or exogenous retroviruses (Thomas and Schneider, 2011).

In a few rare occasions the ERV provirus is fixed in the population because it provided an advantage for the host. As mentioned previously restriction factors such as Fv1, Fv4, enJSRV and others have been co-opted by the host from a retroviral germline invasion, in order to combat other endogenous or exogenous viral infections. However, there are additional examples of ERV co-option for other cellular processes. One of the most striking examples is the recurrent co-option of the fusogenic potential of *env* genes to mediate the required cell-cell

fusion to form the syncytial trophoblast layer during the development of the placenta (Mager and Stoye, 2015). Indeed, different ERV *env* genes have been independently co-opted to perform a similar function in a variety of mammalian orders such as apes, rodents, lagomorphs, ruminants and carnivores (Lavialle et al., 2013), highlighting the major role that ERV co-option has had on the evolution of placentation.

Surprisingly not only protein coding ERVs have been co-opted, but RNA structures and regulatory elements encoded in non-coding regions of the proviral DNA might also provide an advantage for the host to keep. Particularly, the promoter functions of the retroviral LTRs have been redirected to control the expression of specific genes (Diehl et al., 2013; Romanish et al., 2007), but also to control entire regulatory networks (Rebollo et al., 2012), including the mammalian IFN response (Chuong et al., 2016). One particular example is that of MuERV-L, a highly abundant mouse specific endogenous retrovirus that continues to be transcriptionally active at an early stage of the mouse embryo (Kigami et al., 2003; Ribet et al., 2008b). Recent studies have shown that the early embryonic activation of MuERV-L might be in part due to the co-option of its LTRs, whose promoter function has been redirected to control the expression of genes involved in the zygotic genome activation, particularly at the two-cell stage of the mouse embryo (Macfarlan et al., 2011; Macfarlan et al., 2012). Activation of MuERV-L LTRs results in the expression of hundreds of genes that contribute to maintaining the totipotency of the blastomeres, and it also results in the

expression of the MuERV-L Gag-Pol polyprotein and the formation of intracellular viral-like particles (Macfarlan et al., 2011; Macfarlan et al., 2012; Ribet et al., 2008b). In later stages and as development progresses, MuERV-L LTRs and their regulated genes (including MuERV-L) are epigenetically silenced by repressive chromatin modifying enzymes coinciding as well with the expression of pluripotency genes such as Oct4, Sox2 or Nanog (Guallar et al., 2012; Macfarlan et al., 2011; Rowe and Trono, 2011). It is not entirely clear if the presence of these MuERV-L transcripts, proteins and particles at the two-cell stage have an impact on the mouse development, however it is important to point out that the over-expression of MuERV-L during this stage does not correlate with an increase in their copy numbers, suggesting that the expressed proviruses do not have the potential to re-integrate into the genome (Guallar et al., 2012). Additionally, ERVs can also regulate gene expression by the production of long non-coding RNAs (lncRNAs), such as in the case of transcripts promoted by the HERV-H LTR, which seem to play an important regulatory role in human stem cell identity and pluripotency (Fort et al., 2014; Lu et al., 2014).

Paleovirology.

In addition to the previously mentioned co-option events, the discovery of the retroviral “fossil record” also uncovered an exceptional opportunity to expand our knowledge on the co-evolution of retroviruses and their hosts. Previous studies have been able to reconstruct an infectious version of an ancient human

retrovirus by making a consensus sequence of the most recent expansions of HERV-K (HML2) in the human genome (Dewannieux et al., 2006; Lee and Bieniasz, 2007). Those pioneer studies set the groundwork for the new field of paleovirology that is interested in the study of ancient extinct viruses and the effects that such agents have had on the evolution of their hosts (Emerman and Malik, 2010). The HERV-K reconstruction also established that this ancient retrovirus integrates preferentially into transcriptionally active regions, and that members of the APOBEC3 family restricted its replication (Lee and Bieniasz, 2007; Lee et al., 2008). Partial reconstructions of the chimpanzee gammaretroviruses CERV-1 and CERV-2 (present in a variety of primates including old world monkeys but absent in humans) have similarly identified APOBEC3 proteins as strong inhibitors of their replication (Perez-Caballero et al., 2008), although there is evidence that TRIM5α may affect CERV-1 as well (Kaiser et al., 2007). ERV reconstructions have also shed light into the mechanisms of tissue tropism and host range with the identification of copper transport protein 1 (CTR1) as the receptor used by a reconstructed CERV-2 *env* gene and presumably required by the ancient (and possibly extinct) CERV-2 (Soll et al., 2010). Furthermore the same study revealed a set of mutations in CTR1 that made hamster cells resistant to this virus. In parallel, bioinformatic studies identified endogenous lentiviruses in the genomes of leporids (rabbit endogenous lentivirus type K (RELIK)) (Katzourakis et al., 2007; Keckesova et al., 2009), lemurs (prosimian immunodeficiency virus (pSIV)) (Gifford et al., 2008; Gilbert et

al., 2009), ferrets (Cui and Holmes, 2012; Han and Worobey, 2012) and colugos (Han and Worobey, 2015; Hron et al., 2014), contrary to the idea that lentiviruses were a modern retroviral group. Further analysis of RELIK and pSIV allowed reconstruction, expression and crystallization of their Gag proteins, revealing the conservation of the CA-CypA interaction, a feature present in modern day lentiviruses and essential for HIV-1 infectivity (Goldstone et al., 2010). More recently a study followed the evolutionary history of a particular ERV (ERV-Fc) along the entire mammalian class and discovered how exogenous ancestors of this ERV lineage were remarkably able to spread between distant species, highlighting the power of recombination in order to achieve its wide distribution (Diehl et al., 2016).

As previously noted, the existence and sequence of retroviral restriction factors have been shaped by past viral infections. In this regard reconstructions of ancestral TRIM5α proteins have revealed how different ancestral lentiviral groups have shaped lineage-specific variations in modern day TRIM5α proteins (Goldschmidt et al., 2008; McCarthy et al., 2015), as well as modern primate Tetherins (Figure 1.6A). Along the same lines, it is also possible to speculate that ancestral viruses like HERV-K, CERV-1 and CERV-2, might have shaped the evolution of APOBEC3 proteins. Although all these examples illustrate how it has been possible to infer interactions between hosts and viruses, it has not yet been possible to unequivocally demonstrate that specific past retroviral infections were responsible for the origin of any particular restriction factor. Nor has it been

possible to demonstrate that any particular restriction factor was responsible for extinction of any retrovirus (Blanco-Melo et al., 2012). The limited and incomplete nature of the retroviral “fossil” record might difficult answering such questions. However, additional paleovirological resurrections of extinct ERVs might have the potential to discover unidentified restriction factors that were responsible for their extinction and might also act on modern exogenous viruses.

In the following chapters I will describe the paleovirological analysis performed during the course of my Ph.D. work, and I will provide additional insights into the evolution of antique retroviral lineages and the diverse mechanisms responsible for the innovation of host defenses against ancient and modern retroviral infections.

Chapter II. Materials and Methods

Database Integrated Genome Screening (DIGS).

Screening for ERV elements was performed using DIGS (Robert J. Gifford, Unpublished). The DIGS framework is composed of Perl scripts that utilize the BLAST+ package of programs (Camacho et al., 2009). Briefly, a scheme for a simple DIGS screening involves two steps (Figure 2.1): (i) Coding (translated) and non-coding sequences derived from a user defined sequence reference library are used as probes for tBLASTn (translated coding sequences) or BLASTn (non-coding sequences) searches on target genomic sequences. (ii) Significant BLAST hits (e-value < 1E-50) were used as probes for a second round of BLASTx (translated coding hits) or BLASTn (non-coding hits) searches against the previously defined sequence reference library. Significant hits of the first and second round of BLAST searches were added into a relational database to facilitate the management of the screening process and further processing of the results.

An initial reference genome library to screen for ERV sequences was composed of 119 sequences from previously characterized endogenous and exogenous retroviral families. Exogenous retroviral sequences were retrieved from the RefSeq database (Pruitt et al., 2014), and ERV sequences were based either on consensus sequences of previously published ERV sequence data (Benit et al., 2001; Sverdlov, 2005; Tristem, 2000; Villesen et al., 2004), lineage specific mammalian ERVs previously characterized by Robert J. Gifford (unpublished), or

previously inferred consensus sequence analyses (Jern et al., 2005; Lee and Bieniasz, 2007). RT sequences of entire reference genome library were used as probes for the initial “phylogenetic screening” performed by Robert J. Gifford. The probes used for the refinement of the HERV-T and other HERV lineages corresponded of inferred consensus sequences of RT containing elements flanked by paired repeats (likely corresponding to LTRs) identified in the initial “phylogenetic” ERV screen. These sequences were used as well to update the reference sequence library.

A genome target database was built with complete and low coverage primate genome sequences that were retrieved from publicly available databases. For this particular study we selected six ape genomes corresponding to the catarrhini parvorder of higher primates (old world monkeys and apes): human (*Homo sapiens*) (Lander et al., 2001), chimpanzee (*Pan troglodytes*) (Consortium, 2005), bonobo (*Pan paniscus*) (Prufer et al., 2012), gorilla (*Gorilla gorilla*) (Scally et al., 2012), orangutan (*Pongo pygmaeus*) (Locke et al., 2011), gibbon (*Nomascus leucogenys*) (Carbone et al., 2014), baboon (*Papio hamadryas*) (Pruitt et al., 2014), and rhesus macaque (*Macaca mulatta*) (Rhesus Macaque Genome et al., 2007).

The reference sequence library used to screen for MuERV-L elements in the mouse genome contained sequences of endogenous and exogenous Class III retroviral sequences (Spumaviruses, HERV-S, and other ERV-L elements including HERV-L and MuERV-L) retrieved from the general reference sequence

library described above. Amino acid and nucleotide sequences of the MuERV-L reference sequence (MuERV-Lref, GenBank: Y12713) (Benit et al., 1997) were used as probes. Two mouse genome assemblies were used as target genomes: Mm_Celera (NCBI: GCF_000002165.2) (Mural et al., 2002) and GRCm38/mm10 (UCSC: mm10) (Mouse Genome Sequencing et al., 2002).

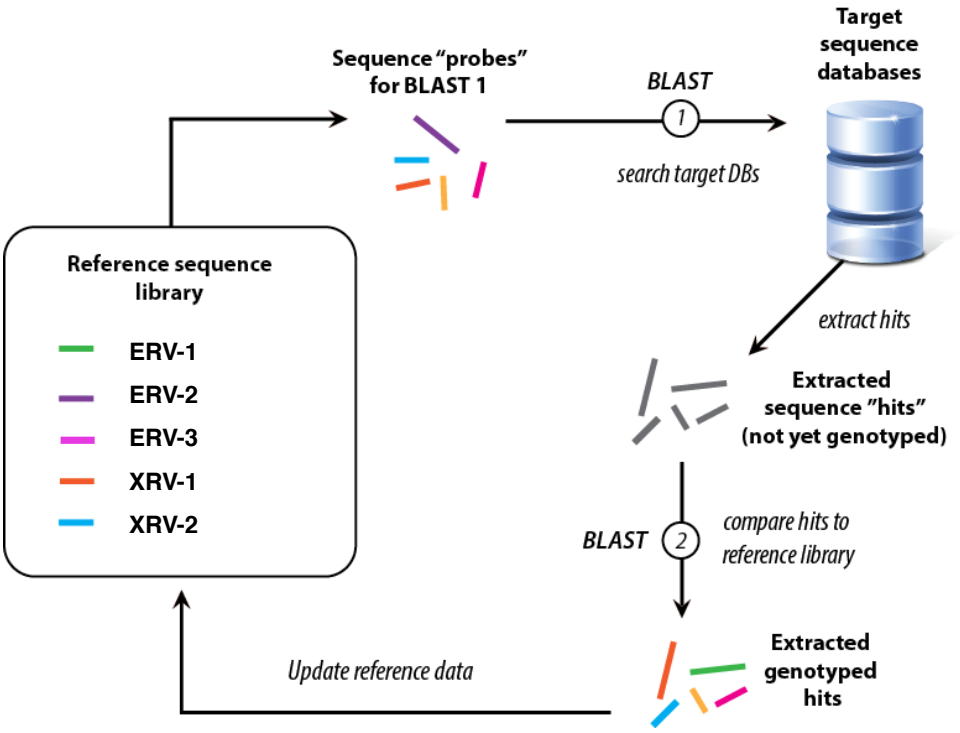


Figure 2.1: Schematic representation of a basic DIGS screening process implemented for ERV discovery.

(1) Sequence probes derived from a user-defined set of endogenous and exogenous reference retroviral sequences are used to screen target databases. **(2)** The reference sequence library is used to classify, or "genotype", significant sequence hits derived from (1). Results might be used to refine the reference library through an iterative process. XRV: exogenous retrovirus. This figure was kindly provided by Dr. Robert Gifford.

Consolidation of DIGS hits.

The consolidation process involves the assembly of adjacent or overlapping hits from the resulting DIGS screening pipeline into a proviral loci. The consolidation process is composed of a series of Perl scripts that retrieve the significant hits (e-values < 1E-50) from the second BLAST search of DIGS, orders them by genome scaffold and orientation and evaluate if two consecutive hits should be "consolidated" following simple rules (Figure 2.2). Two consecutive hits are referred as "consolidated" when the program merges their corresponding coordinates. Briefly, when corresponding to the same genomic feature (same coding or non coding region, i.e. LTR, *gag*, *pol*, *env* or leader), the coordinates of a pair of consecutive hits in the genome scaffold and in the correspondent reference sequence are compared between them and decided to be consolidated following rules depicted in Figures 2.2 A–D. Alternatively, when the pair of consecutive hits involve two distinct genomic features (as defined above) the consolidated process will evaluate their consolidation by keeping the continuous and discrete structure of a provirus (LTR-leader-gag-pro-pol-env-LTR) (Figure 2.2E). The consolidation process allows for insertions no longer than the *length_threshold* parameter (used with the default value of 10,000 nucleotides). Once two consecutive hits are consolidated, the program will treat this consolidated hit as a single hit and will evaluate further consolidation by comparing it to the following one as two consecutive hits (using the process described above). If two consecutive hits should not be consolidated, the 5' most

hit (in the context of its position in the genome scaffold), previously consolidated or not, will be stored in a relational database as a provirus and the program will continue the consolidation of the 3' most hit by comparing it to the following one as two consecutive hits (using the process described above).

Figure 2.2: Rules governing the consolidation of the DIGS results.

Hits corresponding to the same gene: **(A)** two consecutive non-overlapping hits will be consolidated only if the distance between them on the scaffold is less than the *length_threshold*. **(B)** Two consecutive hits with overlapping coordinates on a scaffold will be consolidated. **(C)** Two consecutive hits with overlapping coordinates on a reference sequence will be consolidated only if the distance between them on the scaffold is less than *length_threshold*. **(D)** Two consecutive overlapping hits on both scaffold and reference sequence will be consolidated if the terminal coordinate on the reference sequence of the first hit is smaller than the corresponding coordinate on the second hit. Hits corresponding to different genes: **(E)** consecutive hits will be consolidated into a locus only if the distance between them on the scaffold is less than *length_threshold*, and the assigned gene of the 3' most hit correspond to a succeeding gene on the reference genome structure (LTR-leader-*gag-pro-pol-env*-LTR). All consolidations performed for this work used a *length_threshold* = 10,000nt.

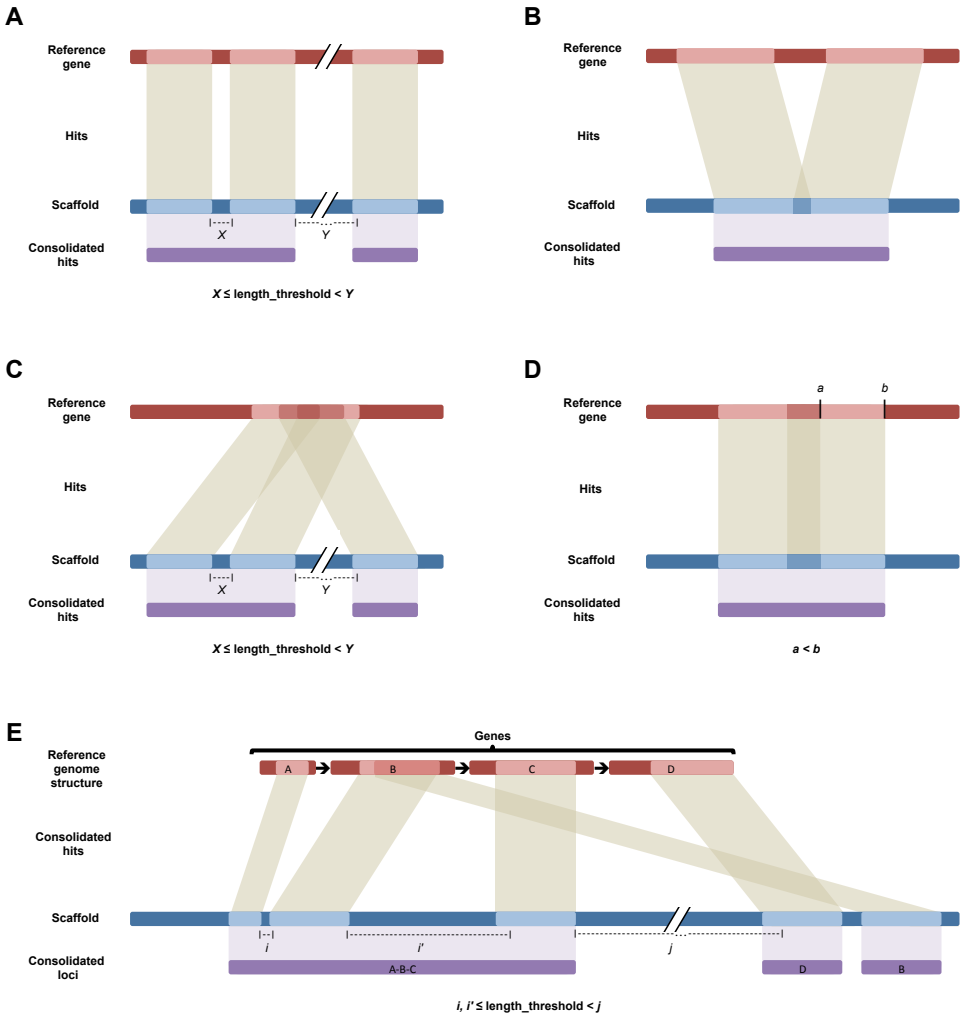


Figure 2.2: Rules governing the consolidation of the DIGS results.

The consolidation process was validated using a synthetic sequence data set. A different Perl script was developed to produce 369 synthetic proviral structures representing the major categories of ERV loci found in animal genomes (i.e. solo LTR, proviruses with paired LTRs, and fragmented loci containing internal regions but lacking paired LTRs). The script takes each artificial proviral structure and assigns it to one of three artificial ERV lineages and inserts it into a random position on one of three artificial scaffolds from each of three artificial organisms (369 loci from 3 artificial ERV lineages distributed between 3 artificial organisms, each of which has 3 artificial scaffolds). Each of the artificial proviral structures was further subjected to a function that randomly "mutates" the proviral structures by introducing insertions or deletions of different lengths. These "mutational" processes also produce distinct small local duplications (resulting in overlapping coordinates) at both the scaffold position (nucleotides) and/or the position in the reference sequence (amino acids). The resulting artificial data set contained 1,132 artificial hits that were used to populate an artificial DIGS result table that was further consolidated into proviral loci by testing different values for the *length_threshold* parameter. The consolidation of synthetic hits into larger loci corresponded exactly with the expected outcome (i.e. the consolidation process was able to revert the random mutations introduced and the initial 369 synthetic proviruses were regenerated).

Multiple sequence alignment of ERV sequences.

Due to the high degree of mutations, especially indels, accurate multiple sequence alignments (MSA) of ERV sequences retrieved from genomic data is challenging. To overcome this difficulty we aligned each retrieved ERV sequence to its corresponding retroviral reference genome sequence individually using MUSCLE (Edgar, 2004), followed by the creation of a "gapped" MSA using the profile alignment function of MUSCLE. Insertions relative to the retroviral reference genome sequence were treated as putative insertions and were eliminated from the MSA, but saved in a separate file. A function that retrieves consolidated DIGS results and creates the MSA as described was programed in Perl. If needed, the same function can reconstruct the "gapped" MSA by taking both the final MSA and the saved putative insertions relative to the retroviral reference genome.

Phylogenetic tree construction.

Maximum likelihood (ML) phylogenetic trees were constructed from a nucleotide MSA using raxML (Stamatakis, 2006) with the following parameters: rapid bootstrap analysis with 10,000 replicates under the GTRCAT model followed by a ML search under GTRGAMMA model to evaluate the final tree topology (-m GTRCAT -# 10000 -x 13 -k -f a). Quick Neighbor Joining trees were constructed

from protein MSA based on distances using PAUP* (set criterion=distance) (Swofford, 2002). Phylogenetic trees were then analyzed using FigTree v1.4.2 (Rambaut, 2008). Thereafter the tree was rooted on the tree midpoint or by an outgroup sequence.

Integration dates of ERV data.

Dates of integration for ERV elements were calculated using PAUP* (Swofford, 2002) by determining the divergence (*K*) to: (i) a consensus sequence, for soloLTRs; (ii) to its cognate LTR, for proviral loci flanked by paired LTRs; (iii) or to the corresponding orthologous sequence, in the case of the complete protein-coding HERV-T *env* locus. The resulting *K* was divided by 2 times the neutral substitution rate (*r*) in order to obtain the estimated integration date (*K*/2*r*) (Lebedev et al., 2000; Subramanian et al., 2011). Neutral substitution rates used for human and mouse were 2.2×10^{-9} and 4.5×10^{-9} substitutions per site per year, respectively (Lander et al., 2001; Mouse Genome Sequencing et al., 2002).

Hypermutation analysis.

Hypermutation analysis and statistics were performed using Hypermut 2.0 (Rose and Korber, 2000) on (i) the set of 230 *gag-pol* containing sequences used to reconstruct the ancestral MuERV-L *gag*, or on (ii) 49 *gag-pol-env* HERV-T2

containing sequences in catarrhine genomes using default parameters or excluding sites with a 5' C next to the mutated G.

Annotation and Statistics of MuERV-L elements in mouse genomic features.

MuERV-L loci were annotated into genomic features for GRCm38 (downloaded using BioMart from Ensembl (Hubbard et al., 2002)) by comparison of their coordinates using in-house Perl scripts. Contingency table and goodness-of-fit tests for MuERV-L elements in the mouse genome were performed using the Pearson's Chi-squared test for count data (chisq.test), implemented in R. Two sets of random coordinates in the mouse genome were computationally generated using in-house Perl scripts and the runif function implemented in R. For comparisons of soloLTRs distributions 10,000 random mouse sequences of 500 nucleotides in length were generated. For comparisons of proviral loci distributions 5,000 random mouse sequences of 6,500 nucleotides in length were generated. The distribution of both random control sets into mouse genomic features was determined as previously described for MuERV-L elements. A third set of 1,000 EcoRV containing random Chinese Hamster sequences of 1,000 nucleotides in length was generated using in-house perl scripts. For ancML integration comparisons each ancML integration site was matched with three random genomic sequences (from the third random control set) that were equidistant to the EcoRV site where the adaptor was ligated. Statistical

comparisons of integration data were performed as described in (Marshall et al., 2007).

Ancestral Reconstructions.

All ML ancestral reconstructions were guided by a MSA together with a phylogenetic tree using baseml from the PAML package (Yang, 1997) (model: REV, initial values of alpha and kappa were calculated on the MSA by jmodeltest (Darriba et al., 2012), branch lengths were used as initial values). A correction for the effect of methylation-induced mutations at CpG islands was applied on both strands of all ancestral reconstructed sequences as described in (Goldstone et al., 2010).

The ancestral reconstruction of MuERV-L was performed on two distinct sequence sets. For the ancestral reconstruction of the *pol* ORF and the LTRs we constructed a MSA and a midpoint rooted ML phylogenetic tree using a set of 95 complete proviral sequences (LTR-*gag*-*pol*-LTR) identified by default BLASTn searches on GRCm38 that were highly similar to MuERV-Lref. Insertions relative to MuERV-Lref in the MSA were treated as putative insertions and were eliminated, except for a 6nt insertion at position 298 and 6249 on both LTRs that were shared between 25% of the sequences. For the ancestral reconstruction of the *gag* ORF we constructed a MSA and a midpoint rooted ML phylogenetic tree on 230 *gag-pol* containing sequences (identified by our previous screening of the mouse genome). We determined the presence or absence of the 33 and 39nt

deletions in the *gag* ORF relative to MuERV-Lref (that does not show any deletion) and identified a monophyletic clade of 40 sequences that showed the 33nt deletion in gag (irrespective of the status of the 39nt deletion). The resulting ancestral *gag*, *pol* and LTR sequences were combined together to produce ancML.

For the ancestral reconstruction on HERV-T we first analyzed the 5' and 3' flanking sequences of 32 HERV-T1 and HERV-T3 proviruses retrieved from catarrhine genomes. These analyses resulted in two and fourteen unique HERV-T1 and HERV-T3 integration events (orthologous groups), respectively. Ancestral nodes for each orthologous group were used to construct the corresponding MSA and ML phylogenetic tree required for the ancestral reconstruction. The HERV-T3 *pre-gag* ORF ancestral sequence was further refined to include the last 43 codons of the *pre-gag* ancestral HERV-T1 sequence and the state of particularly polymorphic sites was hand curated by a combination of their frequency and the phylogenetic relationships between the corresponding sequences. The variation present in HERV-T1 sequences was also taken into account to break possible ties. Logo plots were constructed to visualize the variation present in HERV-T3 and HERV-T1 sequences using WebLogo (Crooks et al., 2004). The final revised sequence corresponded to HTpG. A similar procedure was performed on the initial HERV-T3 ancestral *env* sequence with the addition that if the CpG reversion on a particular position results in the change in chemical nature of the amino acid, the residue at this position is reverted back. The final revised

sequence corresponded to ancHTenv. The ancestral reconstruction for the most recent ancestor of the protein-coding HERV-T3 *env* locus was performed manually by selecting residues based on the phylogenetic relationships of human, gorilla and orangutan. Particularly polymorphic sites were resolved by comparison to the ancHTenv sequence. The final revised sequence corresponded to ancHsaHTenv.

Ancestral PV1 and TM-CC(aT) sequences were inferred on MSA and ML phylogenetic trees constructed from amniote species without any CpG reversions.

In virtro simulation of neutral evolution.

Monte-Carlo simulations of *in silico* neutral evolution on the ancHsaHTenv were performed using seq-gen (Rambaut and Grassly, 1997) under the GTR model (10,000 iterations) as performed in (Katzourakis and Gifford, 2010). Expected branch lengths were calculated using the neutral substitution for human (defined above) for the minimum and maximum estimates of the origin of hominids (13.45 and 19.68 MY respectively) (Perelman et al., 2011). The simulated 10,000 human sequences were then evaluated for the presence of a 5' methionine and the number of stop codons.

mVISTA plots.

Analysis of introns and neutrally evolving sequences of divergent *pv1* and *tm-cc(at)* genes, as well as HERV-T orthologous loci, were carried out using the mVISTA visualization tool for glocal sequence alignments (Frazer et al., 2004). FASTA and annotated sequence files were uploaded to the mVISTA website (http://genome.lbl.gov/vista/mvista/about.shtml) and aligned by the Shuffle-LAGAN algorithm to detect rearrangements (Brudno et al., 2003). A VISTA-Point alignment was generated using a non-overlapping sliding window of 100 bp, and a screenshot of all the alignments was captured.

Identification of TM-CC proteins.

All 105,281 human and 54,447 mouse annotated protein sequences were downloaded from Ensembl (Release 71) using BioMart (Flicek et al., 2013). A Perl script was written to automate a pipeline to find protein sequences that contain one TM domain followed by coiled-coil domains. Protein sequences were scanned for TM domains using tmhmm 2.0 (Krogh et al., 2001). Sequences with only one TM domain and a probability > 0.5 were selected and scanned for coil-coiled segments C-terminal to the TM domain using COILS (probability > 0.5) (Lupas et al., 1991). A complete list of all human and mouse gene locations was retrieved from Ensembl (Release 71) using BioMart. A Perl script was written to compare the locations of the resulting TM-CC proteins and the entire gene location list to establish the identity of the adjacent genes. Homology between

TM-CC proteins that are adjacent to each other was established by BLASTp searches. Similar structural analyses were performed on genomic sequence spanning GTPBP3 and CILP2 (were a contiguous locus was found), Gnomon gene predictions (Souvorov et al., 2010), proteomic and transcriptomic data (Benson et al., 2009) from different animal genomes used in this study. GPI anchor prediction was performed using Pred-GPI (Pierleoni et al., 2008). Splicing variants and expression of a particular TM-CC protein was addressed by consulting public RNAseq data when available.

Analysis of TM-CC(aT), PV1, Tetherin and HERV-T env sequences.

ML analyses on the type of evolution showed by *pv1*, *tetherin* and *tm-cc(at)* genes were performed using CODEML from the PAML suite of programs (Yang, 1997). Due to the high divergence of the N- and C- termini of *tetherin* sequences across mammalian species, the following analyses were performed on a reduced sequence alignment that is composed of the transmembrane (TM) and coiled-coil domains (codons 22 to 133 of human Tetherin). Likewise the PV1 alignment does not include the last two codons (codons 1 to 440 of human PV1). Likelihood ratio tests were used to compare paired, nested models of sequence evolution that allow, or do not allow, variable selective pressure among sites (NSsites models 0 vs. 3), or positive selection (NSsites models 1 vs. 2 and 7 vs 8). The F3X4 model of codon frequencies was used for all analyses in CODEML. The gene trees used in these analyses were constructed using RaxML as previously

described. Model M0 was first used to obtain the branch lengths of the gene trees. The branch lengths were then used as initial values when running more complex models. The chi-square test was performed using 4 degrees of freedom for M0 vs. M3 or 2 degrees of freedom for M1 vs. M2 and M7 vs. M8. For all alignments used, sites where only one or two taxa contain data (i.e. the vast majority of the sites contain gaps) were removed from the alignment.

BLASTp sequence similarity searches were performed under default parameters using either extant or ancestral PV1, TM-CC(aT) or Tetherin protein sequences as queries. The target database contained all Tetherin, PV1 and TM-CC(aT) protein sequences utilized/inferred in this study and the protein products from the 211 human TM-CC genes. A heatmap graph of the e-values associated with the BLAST hits was constructed using Heatmap from Los Alamos HIV databases (http://www.hiv.lanl.gov). Alignment of Coelacanth PV1 and a TM-CC(aT) HMM model was obtained by using the HMMER3 web server (http://www.ebi.ac.uk/Tools/hmmer/) (Finn et al., 2011) with an alignment of TM-CC(aT) protein sequences as query.

Analysis for the features of HERV-T envelope sequences was performed using tmhmm 2.0 (Krogh et al., 2001) for TM and hydrophobic domains, and ProP1.0 (Duckert et al., 2004) for signal peptide and propeptide cleavage sites.

Reconstruction of the tetherin locus.

Sequences proximal to the *tetherin* gene from various animal genomes were retrieved using the Genome Browser at the University of California at Santa Cruz (http://genome.ucsc.edu/index.html) and synteny was assessed using the Genomicus Browser (http://www.genomicus.biologie.ens.fr/genomicus-70.01/cgi-bin/search.pl) and the NCBI Sequence viewer (https://www.ncbi.nlm.nih.gov/tools/sviewer/). In the cases where *pv1*, *tetherin* or *tm-cc(at)* were not annotated in Ensembl, NCBI or UCSC, those genes were identified using orthologous protein queries from closely related species in BLAT searches (Kent, 2002). Publicly available RNA-seq data associated with the *tetherin* locus was analyzed using Ensembl and the NCBI Sequence Viewer.

Plasmid construction.

For the ancML construct we substituted the U3 region of the 5' LTR with a CMV promoter sequence until its TATA box. We also added 12nt containing two MluI sites after the *pol* stop codon to facilitate the cloning of a reporter gene. The modified ancML sequence was synthesized and cloned on to pUC57 expression vector (Genewiz, NJ). Dr. John V. Moran kindly provided the replication dependent L1.3 plasmid (Moran et al., 1999). The replication dependent neomycin resistance (*neo*) cassette (a *neo* gene controlled by a separate SV40 promoter and interrupted by an intron) was PCR amplified from L1.3 plasmid and cloned into ancML using the MluI (New England Biolabs) restriction sites

introduced after the stop codon of *pol*. A separate pCR3.1 expressing GFP and a *neo* gene (NEO) was utilized as a control.

The ancML-RTmut construct was created by overlapping PCR to ancML utilizing primers that annealed to the RT active site with four nucleotide mismatches, resulting in a mutated RT active site (YIDD to AIAA). The PCR reactions were further treated with DpnI (New England Biolabs) restriction enzyme for an hour at 37°C to eliminate plasmid DNA. The complete PCR fragment was cloned back to ancML using unique surrounding BstZ17I and NheI (New England Biolabs) restriction sites contained in the outmost primers.

The ancMLΔGAAGT construct was generated using a reverse primer to the beginning of the PBS and the end of the U5 region of the 5'LTR and that lacked the 5nt linker sequence (GAAGT). The PCR product was DpnI (New England Biolabs) treated for an hour at 37°C to eliminate plasmid DNA. The PCR fragment was cloned back to ancML using unique AgeI and KpnI (New England Biolabs) restriction sites contained in the forward and reverse primers, respectively.

Mouse MOV10 and SAMHD1 (isoform 2) were PCR amplified from total RNA extracted from NIH3T3 cells and cloned into pCR-Blunt II-TOPO using the Zero Blunt TOPO PCR cloning Kit (Life technologies). A plasmid expressing mouse APOBEC3 (C57BL/6J strain) was kindly provided by Dr. Rachel Liberatore (unpublished). HA-tags were introduced into the C-termini of SAMHD1 and APOBEC3 and into the N-terminus of MOV10 by PCR using primers containing

two HA tags and a 15nt linker sequence. All three constructs were introduced into the retroviral expression plasmid pLBCX (Clontech) using unique SfiI (New England Biolabs) sites.

The codon-optimized sequences for expression in human of HTpG, ancHTenv, HsaHTenv and ancHsaHTenv were synthesized (Genewiz, NJ) and subsequently cloned into the pCAGGS expression vector (Niwa et al., 1991) using EcoRI and XhoI enzymes. Furin site modified HERV-T envelopes were generated by interchanging the active furin cleave site of ancHTenv for the inactive one in HsaHTenv, and *vice versa*. Primers that annealed to the respective furin cleavage sites with eigth nucleotide mismatches were used in overlapping PCR reactions, resulting in the interchange of the furin cleavage site to create ancHTenv-FurinMut (SRFRRAA to PRLHQAV) or HsaHTenv-FurinFix and ancHsaHTenv-FurinFix (PRLH(QIR)AV to SRFRRAA). The PCR reactions were further treated with DpnI (New England Biolabs) restriction enzyme for an hour at 37°C to eliminate plasmid DNA. The complete PCR fragment was cloned back into pCAGGS using EcoRI and XhoI (New England Biolabs) restriction sites contained in the outmost primers. HA tagged versions were produced by introducing two HA tags into the C-termini of all HERV-T envelopes, and into the N-terminus of HTpG, by PCR using primers containing two HA tags, a 15nt linker sequence and at least 30nt annealing to the corresponding termini. Tagged and un-tagged HERV-T envelope constructs were subcloned into pCCIB and pCCIGW (Kane et al., in preparation) lentiviral expression plasmids to produce

stable cell lines and for transient expression (for FACS experiments), respectively. HERV-T envelope sequences and SIV Nef were PCR amplified using primers containing SfiI (pCCIB) or SnaBI and BstXI (pCCIGW) for subcloning. The sequence of human MOT1 was amplified from the gDNA of DF-1 cells expressing a 293T cDNA library (described below). The amplified sequence was cloned into pCCIB utilizing SfiI sites to make stable DF-1 cell lines expressing human MOT1. Human SERINC5 construct was kindly provided by Dr. Fengwen Zhang (unpublished). SIV and HIV Nef constructs have been previously described (Zhang et al., 2009). The MLV glycoGag construct was kindly provided by Dr. David Perez-Caballero (unpublished). The MLV Gag-Pol (MLVgp), MLV-A and MLV-E expressing plasmids were kindly provided by Dr. François-Loïc Cosset.

All Tetherin, PV1, TM-CC(aT) or TM-CC proteins were transiently expressed using pCR3.1 (Invitrogen) based plasmids. The human PV1, the Tasmanian devil Tetherin, human CD72 and mouse CLEC1A expression plasmids were kindly provided by Dr. Siddarth Venkatesh (unpublished). All remaining *tetherin/tm-cc-gpi* (except human) and *tm-cc(at)* genes were synthesized with codon-optimization for human cells (Genewiz, Inc.) and cloned into pCR3.1 using EcoRI and NotI restriction sites.

The sequences used in this study were: (i) Tetherin/TM-CC-GPI proteins: Human (NP_004326.1), Opossum (XP_007489270.1), Tasmanian devil (XP_012399618.1), Chinese alligator (XP_006017476.1), and Elephant shark

(XP_007897024.1). The Falcon Tetherin protein was derived from the incorrect gene annotation of the CILP2 gene (XM_005444350.1), which resulted in a fusion of CILP2 and Tetherin and was guided by Gnomon prediction 2189215010.p. The Coelacanth Tetherin protein was derived from a Gnomon gene prediction (16424589.p) at the genomic sequence NW_005819727.1 from positions 98723 – 108999. The lamprey TM-CC-GPI sequence was derived from the TM-CC-GPI analysis (see above) of assembled transciptome data from the Lamprey sequencing consortium (Smith et al., 2013) (transcript number 1626467). (ii) TM-CC(aT) proteins: Mouse (XP_003945491.1), Painted turtle (XP_008169839.1). Human TM-CC(aT) was derived from GenBank entry XP_011526778.1 excluding the predicted fifth exon so that it has four exons similar to the mouse variant. Chinese alligator TM-CC(aT) was derived from a Gnomon gene prediction (2147496003.p) and refined using the predicted American alligator TM-CC(aT) protein (KQL90195.1). Coelacanth TM-CC(aT)$_B$ was derived from GenBank entry XP_006001674.1 with the first 12 amino acids removed, guided by an alignment of other TM-CC(aT) proteins. (iii) PV1 protein: Human (NP_112600.1).

All GPI anchor additions were generated by using a reverse primer that encoded the Tetherin GPI anchor sequence (amino acids 160 to 180) and a NotI site. HA-tagged versions of all the constructs were generated using a forward primer encoding 2 repeats of the HA tag previous to the initiation codon. Alternatively

spliced variants of TM-CC(aT) were constructed by PCR using different 3' primers.

The sequence for all plasmids was verified by Sanger sequencing (Genewiz, Macrogen).

Cell culture.

All cells used in this study (except CHO-K1 and pgsA cells) were maintained in Dulbecco's Modified Eagle Medium (DMEM), Eagle's Minimum Essential Medium (EMEM) or Roswell Park Memorial Institute medium (RPMI) supplemented with 10% FBS and gentamycin (2μg/ml, Gibco) according to ATCC instructions. CHO-K1 and pgsA cells were maintained in Ham's F-12 media supplemented with 10% FBS, 1mM of L-glutamine and 2ug/ml of gentamycin. All cells were incubated at 37°C, except DF-1 cells that were incubated at 39°C.

In order to make stable cell lines, 293T cells were transfected with plasmids expressing MLV gag-pol polyprotein (MLVgp), VSV glycoprotein (VSV-g) and pLBCX plasmids containing HA tagged or untagged versions of mouse APOBEC3, MOV10, SAMHD1 (described above) using polyethylenimine. Alternatively, 293T cells were transfected with plasmids expressing HIV-1 gag-pol polyprotein (pCRV1, (Zennou et al., 2004)), VSV glycoprotein and pCCIB plasmids containing human MOT1 or HA tagged/untagged, mutated/non-mutated versions of different HERV-T envelopes (described above) using

polyethylenimine. In every case viruses were harvested and filtered two days after transfection and were used to infect the naïve cells of interest (seeded in 24 well plates). Infected cells were expanded in 10 cm dishes with media supplemented with the corresponding amounts of blasticidin S (Thermo Fisher Scientific Inc.) and were monitored from 3 to 10 days before performing experiments or isolating single cell clones.

Single cell clones expressing mouse APOBEC3, MOV10 and SAMHD1, or different HA tagged/untagged, mutated/non-mutated versions of HERV-T envelopes were further generated by expanding blasticidin resistant cells seeded at 0.5 cells per well in a 96 well plate. Single colonies on 96 well plates were monitored and sequentially expanded.

MuERV-L replication assays.

Sixteen cell lines (Table 4.1) were seeded on 12 well plates one day before being transfected with 700ng of plasmids expressing L1.3, ancML or a plasmid expressing *gfp* and a *neo* gene using 4μl of Lipofectamine 2000 (Thermo Fisher Scientific Inc.) according to manufacturer instructions. DNA and Lipofectamine 2000 dilutions were performed using Opti-MEM (Gibco). Cells were expanded into 6 well plates under G418 selection media two days after transfection. Amounts of G418 were titrated previously for each cell type. Ten days after selection cells were fixed with 4% paraformaldehyde (PFA) and colonies were stained using 0.3% crystal violet in 20% ethanol for counting.

The ancML replication assays on CHO-K1 cells were performed as follows. CHO-K1 cells were seeded at a concentration of 3 x 10^5 per well on a 12 well plate. One day after, the cells were transfected with 1μg of plasmid DNA using 3μl of Transit-CHO supplemented with 0.5μl of CHO-mojo reagent (Mirus) diluted in Opti-MEM (Gibco). One day after, the cells were expanded on a 10cm dish with media supplemented with or without different concentrations of Zidovudine (AZT) or mouse INF-α (Pestka Biomedical Laboratories, Inc.). Two days after, cells were expanded on a 15 cm dish or three 96 well plates (for analysis of single cell clones) with media supplemented with 1μg/ml of G418. 1/1000 of the cells initially transfected with the NEO plasmid were expanded on a 15 cm dish with selection media. Cells in 15 cm dishes were cultured under selection for an additional 10 days before treatment with 4% PFA and colonies were stained with 0.3% crystal violet in 20% ethanol for counting. Single colonies on 96 well plates were monitored and sequentially expanded until reaching confluency on a 10 cm dish. Genomic DNA (gDNA) was extracted of 5 x 10^6 cells using QIAmp DNA mini kit (QIAGEN) for genome walker assays (see below). AZT was obtained from the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH.

Integration sites analyses (Genome Walker).

Integration sites of ancML were cloned and sequenced using the universal genome walker kit (Clontech). Nested PCRs were performed on gDNA of single cell clones of CHO cells transfected with a plasmid expressing ancML and

previously digested with EcoRV (New England Biolabs) and ligated to adaptors. Forward primers to amplify 3' flanking sequences were designed to anneal to the R region of the 3'LTR and the reverse primers to the adaptor sequence. Reverse primers to amplify 5' flanking sequences were designed to anneal to the U3 region of the 5'LTR and the forward primers to the adaptor sequence. Major bands from secondary PCRs were gel purified and cloned into pCR-Blunt II-TOPO using the Zero Blunt TOPO PCR cloning Kit (Life technologies) for sequencing. Resulting CHO gDNA sequences were mapped to the CHO genome (criGri1) using BLAT (Kent, 2002) searches on the UCSC genome browser (Kent et al., 2002). Similar searches were performed on the matched random sequences.

PCR analyses.

To estimate the fate of the intron interrupting the *neo* gene, gDNA of CHO cells transfected with a plasmid expressing ancML, ancMLΔGAAGT or an empty vector, were used as template for PCR analysis. Forward and reverse primers were designed to anneal to the extremes of the *neo* gene. Human MOT1 and the non-coding fragment of human atg12 were PCR amplified from gDNA of RFP expressing DF-1 cells infected with a 293T cDNA retroviral library (described below) and resistant to G418 and Hygromycin. Forward and reverse primers were designed to anneal to the extremes of the pCCIB vector in which the cDNA library was cloned into (described below). Mouse TM-CC(aT) was PCR amplified,

cloned and sequenced from cDNA derived from 7-day mouse embryo total RNA (Clontech, Cat. # 636607) using forward primers specific to the first and third exon and reverse primers specific to the third exon and to the 3'UTR. For all PCRs performed in this study we used Phusion High-Fidelity DNA Polymerase (ThermoFisher).

Virion yield assays.

MLV particles pseudotyped with the amphotropic envelope (MLV-A), the ecotropic envelope (MLV-E), VSV-g and different HERV-T envelopes were produced in 293T cells by co-transfecting the corresponding envelope plasmids with plasmids expressing a MLV gag-pol polyprotein (MLVgp), a *neo* gene and GFP/RFP (pCNCG/pCNCG) (Soneoka et al., 1995) or a hygromycin resistance gene (pLHCX, Clontech), using polyethylenimine. Additionally for certain experiments, 293T cells were also co-transfected with increasing amounts of plasmids expressing HTpG, MLV glycoGag, SIV Nef, HIV Nef (pCAGGS), and for some experiments also with 15μg of a plasmid expressing human SERINC5. In every case the total amount of DNA was held constant by supplementing the transfection with an empty vector (pCAGGS). Viruses were harvested and filtered (0.22 μm) two days after transfection and serial dilutions (1/3) were used to infect the cells of interest (293T, DF-1, NIH3T3, HT1080, etc). Viral titers were calculated by expanding the infected cells in 10 cm dishes with media supplemented with the corresponding amounts of antibiotic and monitored for 10

days before resistant colonies were counted (for MLV particles expressing *neo* and hygromycin resistance genes). Alternatively, viral titers were calculated by determining the percentage of infected cells expressing GFP or RFP 2 days post infection using the Guava EasyCyte flow cytometer (Millipore). Where specified, viruses were concentrated using the Amicon Ultra-15 filters (10kDa, Millipore).

HIV-1(WT) and HIV-1(ΔVpu) versions of the HIV-1 molecular clone NL4-3 have been previously described (Neil et al., 2006). 293T cells were co-transfected using polyethylenimine with 300ng of wild-type (HIV-1(WT)) or Vpu-deficient (HIV-1(ΔVpu)) proviral plasmids along with varying amounts of Tetherin, PV1, TM-CC(aT) or TM-CC expression plasmids and a plasmid expressing YFP (100ng), to monitor transfection efficiency. For experiments with Tetherin proteins, 5 to 20 ng of plasmid were used. In all transfection experiments, the total amount of DNA was held constant by supplementing the transfection with an empty expression vector (pCR3.1). Two days post transfection, the culture supernatants were harvested, clarified by centrifugation at 3000 rpm, and filtered (0.22 μm). Infectious virus yield was determined by inoculating sub-confluent monolayers of HeLa-TZMbl cells (NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH) that were seeded in 96 well plates at 10,000 cells/well with 100 μl of serially diluted supernatants. At 48 hours post infection, β-galactosidase activity was determined using GalactoStar reagent, in accordance with the manufacturer's instructions (Applied Biosystems).

For all experiments, physical particle yield was determined by layering 600 μl of the virion containing supernatant onto 1 ml of 20% sucrose in PBS followed by centrifugation at 20,000xg for 90 minutes at 4°C. Virion pellets were then analyzed by Western blotting.

Western blot assays

Pelleted virions and cell lysates were resuspended in SDS-PAGE loading buffer, with the addition of 0.5% β-mercaptoethanol, and resolved on NuPAGE Novex 4-12% Bis-Tris Mini Gels (Invitrogen) in MOPS running buffer. Proteins were blotted onto nitrocellulose membranes (HyBond, GE-Healthcare) in transfer buffer (25 mM Tris, 192 mM glycine). The blots were then blocked with Odyssey blocking buffer and probed with mouse monoclonal anti-HIV-1 capsid (R183, NIH), rat monoclonal anti-MLV capsid (R187, ATCC), mouse monoclonal anti-HA (Covance), or mouse monoclonal anti-PV1 (Abcam) primary antibodies. The bound primary antibodies were detected using a fluorescently labeled secondary antibody (IRDye 800CW Goat Anti-Mouse Secondary Antibody, LI-COR Biosciences). Fluorescent signals were detected using a LI-COR Odyssey scanner and quantitated with Odyssey software (LI-COR Biosciences).

Immunofluorescence assays.

Transfected, stable cell populations or single cells clones expressing a particular HA-tagged HERV-T envelope or HTpG were seeded one day previous to the immunofluorescence assay. Cells were fixed with 4% PFA for 30 minutes followed by treatment with 10mM glycine (diluted in PBS) for another 30 minutes. Cells were permeabilized with a buffer containing 0.1% of Triton X-100 (ThermoFisher) and 5% goat serum (diluted in PBS) for 15 minutes. Cells were then washed 2 times with PBS before being treated with mouse monoclonal anti-HA (Covance) diluted in a buffer containing 0.1% Tween-20 (ThermoFisher) and 5% goat serum (diluted in PBS) for 2 hours at room temperature. Cells were washed three times with PBS before being treated with goat anti-mouse secondary antibody (Alexa Fluor 488 dye, ThermoFisher) diluted in a buffer containing 0.1% Tween-20 (ThermoFisher) and 5% goat serum (diluted in PBS) for 1 hour at room temperature. DNA was stained with 50μg/ml of 4',6-diamidino-2-phenylindole (DAPI) (diluted in PBS) (Invitrogen) for 5 seconds. Cells were washed three more times with PBS and fluorescent microscopy images were analyzed using the DeltaVision software (GE Healthcare) or the EVOS FL Cell Imaging System (ThermoFisher).

Fluorescence-Activated Cell Sorting (FACS).

Viruses were produced in 293T cells transfected with plasmids expressing HIV-1 gag-pol polyprotein (pCRV1), VSV glycoprotein and pCCIGW plasmids

expressing GFP and containing mutated/non-mutated versions of different HERV-T envelopes (described above) or an unrelated protein (SIV Nef) using polyethylenimine. Viruses were harvested and filtered (0.22 μm) two days after transfection and were used to infect 2 x 10^5 naïve 293T cells seeded in 24 well plates. Transduced cells were expanded in 10 cm dishes until reaching confluency. Afterwards, serial dilutions of concentrated frozen stocks of MLV expressing RFP (pCNCR) pseudotyped with ancHTenv were used to infect 2 x 10^5 transduced 293T cells in 24 well plates. Two days post infection the number of GFP positive, RFP positive and double positive cells were counted by FACS using CyFlow space (Partec). The resulting data was analyzed with the FlowJo analysis software. The percentage of RFP and GFP positive cells was calculated by gating on the GFP positive population.

293T cDNA library preparation and screening.

Total RNA was isolated from a confluent 10cm dish of 293T cells using Trizol (Invitrogen). mRNA transcripts were enriched using Oligotex polyA+ resin (Qiagen). polyA RNA was used to construct a cDNA library using the SMART cDNA Library Construction Kit (Clontech). Briefly, cDNA containing SfiI restriction sites was synthesized using the SMARTScribe MMLV reverse transcriptase (Clontech) with SMART primers. The resulting cDNA was further amplified by 15 cycles using Phusion High-Fidelity DNA Polymerase (ThermoFisher) using SMART primers. PCR products were treated with Proteinase K (Clontech) for 20

minutes at 45°C before they were digested by SfiI restriction enzyme (NEB) for 2 hours at 50°C. Digested products were then size fractionated using CHROMA SPIN-400 columns (Clontech). The first cDNA containing fractions (>500bp) were selected to ligate into a previously digested pCCIB plasmid with the corresponding SfiI sites. Overnight ligation was performed using T4 DNA ligase (NEB). The resulting 293T cDNA library had a complexity of 3.5 x 10^6 colony forming units and was expanded by transformation into electrocompetent DH5α bacteria cells. Transformed bacteria were cultured on 6lt of SeaPrep soft agarose (Lonza) diluted in 2xLB media at 30°C for 2.5 days, followed by DNA preparation.

Viral stocks carrying the cDNA library were produced by transfecting 6 x 10^6 293T cells with plasmids expressing HIV-1 gag-pol polyprotein (pCRV1), VSV glycoprotein and the cDNA library plasmids (pCCIB). Viruses were harvested, filtered (0.22 μm), concentrated (Amicon Ultra-15 filters 10kDa, Millipore) and frozen two days after transfection. DF-1 cells (or NIH3T3) were infected with the library at an MOI of 8. Infected DF-1 cells were challenged with frozen stocks of MLV particles pseudotyped with ancHTenv containing a *neo* resistance gene (pCNCG) two days after infection with virus carrying the cDNA library. Infected DF-1 cells were placed in G418 selection two days post infection (dpi) with the *neo* containing virus and resistant colonies were collected after 10 days. G418 resistant DF-1 cells were then challenged with ancHTenv pseudotyped MLV particles containing a hygromycin resistance gene (pLHCX). Cells were placed in hygromycin selection two dpi and resistant colonies were collected after 10 days.

Hygromycin resistant DF-1 cells were infected with ancHTenv pseudotyped MLV particles expressing RFP upon infection (pCNCR). gDNA was extracted from the RFP positive DF-1 cell population and possible receptor candidates were amplified by PCR using primers to the pCCIB vector.

Chapter III. Screening for primate and murine ERVs

Essential for any paleovirological analysis is the initial screening to assess the endogenous viral elements (EVEs) diversity present in the genome(s) of interest. Comparative investigations based on sequence similarity searches have been crucial in the discovery and characterization of EVEs (Katzourakis and Gifford, 2010; Katzourakis et al., 2007; Tristem, 2000). However given the large volumes of sequence data available, an organized similarity-search based approach is needed to identify highly mutated EVEs in multiple genomic sequences.

Database Integrated Genome Screening.

We collaborated with Dr. Robert Gifford at the MRC-University of Glasgow Center for Virus Research, to develop DIGS (database-integrated genome screening), a computational framework to identify EVEs in a variety of animal genomes by implementing systematic BLAST-based *in silico* screens of molecular sequence databases.

The DIGS framework is composed of Perl scripts that utilize the BLAST+ package of programs (Camacho et al., 2009) to identify, extract, and classify EVEs present in genomic sequences. It requires a control file that specifies the parameters used for the BLAST-based screening, as well as a set of target nucleotide sequence databases for screening (i.e. complete genomes or scaffolds), and a reference library containing well characterized viral sequences.

A basic DIGS screening project is comprised of two steps (Figure 2.1). In the first step, query sequences selected from the reference library are used to search the target databases. In the second step, significant hits from the first round of BLAST are classified, or "genotyped", by a second BLAST comparison to the reference library. The screening progress and their results are captured in a relational database enabling both the management of the screening process and the interrogation of the data generated, using structured query language (SQL). Subsequent analysis of the results might be used to refine the reference library allowing the screen to proceed through multiple iterations and reconfigurations.

Dr. Robert Gifford performed an initial screening in which he used RT amino acid sequences of endogenous and exogenous retroviruses as probes to identify ERV sequences in catarrhine genomes (old world monkeys and apes). This initial "phylogenetic" screening exploited the fact that the retroviral RT is relatively resistant to mutation, and can be used for molecular phylogenetic investigations across the entire retroviridae family. Therefore, RT sequences recovered by this screening were placed into an alignment containing RT sequences of well-characterized retroviral families, and used to construct a phylogenetic tree. As expected, all catarrhine ERV RTs were clustered into monophyletic lineages inside three major clades, corresponding to the three major divergences in the evolution of retroviral RTs (Llorens et al., 2008; Tristem, 2000), and grouped outside the clades defined by exogenous retrovirus genera. Eighteen of these

ERV lineages were selected for further investigation, representing four major RT clades (Table 3.1).

We inferred consensus genomes of those 18 ERV lineages by extracting and aligning sequences containing a RT flanked by paired repeats (likely to constitute LTRs). The major polypeptides and non-coding LTR sequences of those consensus genomes were used as probes on a second round of DIGS screening in catarrhine genomes. Individual results of this screening were automatically assembled into representative proviral loci using a "consolidation" function adapted for DIGS. This function orders the DIGS results by genome scaffold and orientation; adjacent or overlapping entries are assembled into a proviral locus by comparison with a canonical ERV reference genome structure of the form LTR-*gag*-*pol*-*env*-LTR (see chapter II and Figure 2.2). The consolidation function allows not only the defragmentation of individual results into a contiguous proviral locus, but also can potentially uncover possible recombination events by consolidating ERV sequences that were classified to belong to distinct ERV lineages. The results of this second round of DIGS screening and consolidation are summarized in Table 3.1.

Table 3.1: Analysis of 18 ERV lineages present in catarrhine genomes identified by DIGS

Lineage	RT Clade	Total	*pol*[#]	soloLTR	Provirus	*env*+	*env*-	Recombinant Loci*
HERV-L	L/S	24919	2033	17698	1140	11	7210	5.28%
HERV-S	L/S	522	190	32	9	206	284	4.29%
ERV-9	9/30/W	22293	379	19197	797	1358	1738	14.34%
HERV-30	9/30/W	416	3	267	33	32	117	60.40%
HERV-W	9/30/W	5446	182	3054	533	409	1983	37.63%
HERV-H	H/XA/F	11430	651	5718	2339	499	5213	8.40%
HERV-XA	H/XA/F	421	57	250	19	53	118	20.47%
HERV-Fb	H/XA/F	547	29	264	25	95	188	11.31%
ERV-Fc	H/XA/F	388	28	235	23	51	102	12.03%
HERV-E	E/T	2771	96	734	127	1181	856	29.11%
HERV-T	E/T	1516	12	1166	127	181	169	20.57%
HERV-K14C	HERV-K	814	70	450	66	126	238	45.33%
HERV-K-HML2	HERV-K	12110	55	11410	167	381	319	23.43%
HERV-K-HML4	HERV-K	2189	3	2062	39	96	31	33.07%
HERV-K-HML5	HERV-K	809	219	105	25	295	409	13.07%
HERV-K-HML6	HERV-K	504	17	99	15	250	155	11.60%
HERV-K-HML8	HERV-K	1535	18	1293	69	199	43	31.82%
HERV-K-HML9	HERV-K	2322	369	788	70	575	959	16.04%

[#]Proviruses classified as *pol* are those that only contain *pol* sequences. * Recombinant loci refer to the percentage of non-soloLTR loci composed of hits belonging to more than one assigned lineage.

As expected, soloLTR sequences are in high excess compared to other proviral structures for most of the lineages analyzed. The low soloLTR proportion for some lineages can be explained by the alternative usage of other related LTR sequences. This is the case of HERV-S where the soloLTRs account for only 6% of all the loci. According to RepBase (Jurka et al., 2005) HERV-S elements can be associated with 3 types of LTR sequences LTR18A, B and C, which share 65 - 79% nucleotide identity between them, but with a high proportion of indels. It is possible that the lack of HERV-S soloLTRs identified in this analysis is due to the underrepresentation of other related LTR sequences in the consensus reference genome and in the reference library. A similar scenario might be true for lineages such as HERV-Fb and HERV-E in which their elements are associated to diverse LTR groups.

All lineages selected contained a percentage of sequences that might have a recombinant origin. For HERV-E members, where the percentage of non-soloLTR recombinant loci is close to 30%, the main contributors are sequences assigned to HERV-W, HERV-I, HERV-K-HML2 and ERV-9. The Harlequin recombinant element is composed of pieces of HERV-E, HERV-W, HERV-I and HERV-P (Jurka et al., 2005), thus it is possible that some of the HERV-E mosaic loci identified by DIGS represent Harlequin elements. Similarly other HERV-E mosaic loci might account for the low copy DA and Xiao elements composed of HERV-E, HERV-K, HERV-T and HERV-H fragments (Ji and Zhao, 2008; Li et al., 2009). Sequences belonging to the ERV-9/HERV-30/HERV-W RT clade showed

a high degree of mosaicism within elements of the same clade, this is particularly the case for HERV-30 where 60% of their non-soloLTR loci are of recombinant origin and are mainly composed of sequences assigned to ERV-9 (52%) and HERV-W (20.81%). This high degree of putative recombination in a RT clade can also be an artifact of the "genotypification" step implemented in DIGS, where closely related sequences can be incorrectly assigned to a member of the same clade. A similar scenario might be true for members of the HERV-K RT clade, where the percentage of recombinant non-soloLTR loci ranges from 11% for HERV-K-HML6, to 45% for HERV-K14C. Of this putative mosaic species, a high proportion contain sequences assigned to other HERV-K members (particularly HERV-K-HML2, 9 and 14) and, to a lower extent, sequences assigned to ERV-9 and HERV-H. The opposite scenario seems to be true for the abundant HERV-H lineage where the majority of its loci are composed solely of HERV-H sequences.

HERV-H possess the most proviruses flanked by two LTRs, 2,339 loci or ~20% of all HERV-H loci, including soloLTRs. Larger lineages are abundant in sequences that had replicated by complementation *in trans* or by retrotransposon-like replication (Figure 1.7B) (Belshaw et al., 2005). Consistent with the inference that they had arisen via these processes, such large lineages overwhelmingly lacked complete ORFs (Magiorkinis et al., 2012) or showed 5' truncations and poly-A tails indicative of LINE retrotransposition (Esnault et al., 2000). In the case of HERV-H there are 10 fold more loci with missing *env* genes than *env* containing loci, which argues for a retrotransposon-like replication, in

agreement with previous observations (Goodchild et al., 1995). A similar scenario is true for HERV-L and HERV-W (Belshaw et al., 2005; Magiorkinis et al., 2012). HERV-L is distantly related to spumaviruses and prototypical members show a complete lack of an *env* gene. In the current study, DIGS was able to characterize 11 proviruses containing an *env* gene, however a closer look into these sequences show mosaic origin with predominantly HERV-L LTRs flanking *env* genes belonging to HERV-E and HERV-I. Overall DIGS successfully identified and assembled loci associated with the selected lineages in catarrhine genomes with numbers in accordance to previous estimates.

Analysis of a human ERV lineage closely related to gammaretroviruses.

Although no extant human gammaretroviruses have been identified, HERV-T is an ERV lineage that shares significant similarity with members of the gammaretrovirus genus (Blusch et al., 1997). Indeed, the initial "phylogenetic" screening found HERV-T as the lineage closer to the gammaretroviruses clade. Our previous DIGS screening revealed relatively low copy numbers in catarrhine genomes with soloLTRs accounting for more than 75% of all elements. HERV-T also has a moderate rate of mosaicism (20%) mostly with lineages outside its RT clade (mainly HERV-H, HERV-K, ERV-9 and HERV-30). Additionally, and consistent with its low copy number, HERV-T has an envelope sequence associated with the majority of their non-soloLTRs elements (52%). Previous analysis of these envelope sequences has revealed a significantly low dN/dS

ratio suggesting past purifying selection and a replication mechanism dependent on reinfection (Belshaw et al., 2005).

In order to further investigate the evolution of this lineage, we downloaded 44 non-mosaic proviral sequences containing coding fragments (*gag*, *pol* and *env*) flanked by paired LTRs and aligned them to the consensus reference HERV-T sequence (see chapter II). The resulting alignment was used to guide a maximum likelihood phylogenetic tree that was rooted based on an out-group sequence, a HERV-T related sequence in the genome of a platyrrhine (new world monkey) primate (*Saimiri sciureus*) (Figure 3.1). The phylogenetic tree clearly identifies three distinct monophyletic clades. This topology was corroborated on phylogenetic trees obtained using an alignment of the soloLTRs. A closer look into the LTRs of HERV-T shows three classes of LTRs, which correspond to the paired LTRs of the three monophyletic clades in Figure 3.1. Two were previously annotated in RepBase, LTR6A and LTR6B (corresponding to HERV-T3 and HERV-T2 in our notation), and a more divergent third one (HERV-T1) that shares 55% and 63% of identity with the other two, respectively (Table 3.2). On average in alignments comparing HERV-T2 or HERV-T3 to HERV-T1 LTRs, 26% of sites correspond to gaps, inserted primordially in the first two thirds of the alignment (Table 3.2). Each of these LTR classes might account for three independent germ line invasion events, or for three expansion periods of a subset of ancestral HERV-T elements. In order to obtain more information about the dynamics of these lineages, and given that there are relatively low numbers of paired LTR

elements (Table 3.1), we extracted each soloLTR identified in catarrhine genomes, classified them to belong to one of the three classes of LTRs and create a consensus sequence of each class. Each LTR was then compared to the corresponding consensus sequence and its divergence was calculated using PAUP* (Swofford, 2002) under a GTR model (see chapter II). Integration dates for each LTR were subsequently calculated by adjusting the distance to the consensus using a human neutral substitution rate of 2.2×10^{-9} substitutions per site per year (Lander et al., 2001; Subramanian et al., 2011) (Figure 3.2). The cumulative lineage through time plots for each HERV-T class indicates that HERV-T3 is the oldest. HERV-T3 started a slow expansion before the old world monkeys diverged from the rest of the apes, and might have derived from the HERV-T-like elements present in platyrrhini species. HERV-T3 elements then started a fast expansion period between ~25 and 15 MYA, probably through reinfection mechanisms, as indicated by the integrity of the envelope and further analysis described in chapter V. HERV-T1 slowly expanded during ~15 MY and showed a boost about 15 MYA. HERV-T2 might have derived from a HERV-T1 (Figures 3.1 and 3.2) and showed an initial slow expansion with a further boost ~20 MYA, about the times gibbons diverged from the rest of the hominids. These estimated dates of integration were further corroborated with dates calculated from paired LTRs containing elements (see Chapter V).

Table 3.2: Percentage of identity and gaps inserted in pairwise alignments of HERV-T LTR classes.

LTR class	HERV-T1		HERV-T2		HERV-T3	
	% Identity	% Gaps	% Identity	% Gaps	% Identity	% Gaps
HERV-T1	100	0				
HERV-T2	63.6	24.2	100	0		
HERV-T3	55.1	28.9	67.7	17.8	100	0

Alignments generated on consensus LTRs using a Needleman-Wunsch algorithm.

As observed in the phylogenetic tree of HERV-T proviruses (Figure 3.1), a subgroup of HERV-T2 sequences clustered into a tight monophyletic clade of closely related sequences, depicted by short branch lengths compared with the rest of HERV-T sequences. A close examination of all non-mosaic HERV-T2 proviral elements, containing fragments of coding genes (*gag-pol-env*), revealed that 59% of these elements showed significant signatures of APOBEC3-mediated hypermutation (Rose and Korber, 2000) (Figure 3.3), and share similar sequence deletions compared to a consensus HERV-T sequence, particularly involving *gag* and *env* genes. These observations suggest that the majority of the copy numbers observed for HERV-T2 elements were derived from a reduced set of hypermutated and inactivated elements (most likely due to APOBEC3 proteins) that were expanded by complementation *in trans*, probably through retrotransposon-like mechanisms. Further paleovirological analyses were performed on HERV-T1 and HERV-T3 elements described in chapter V.

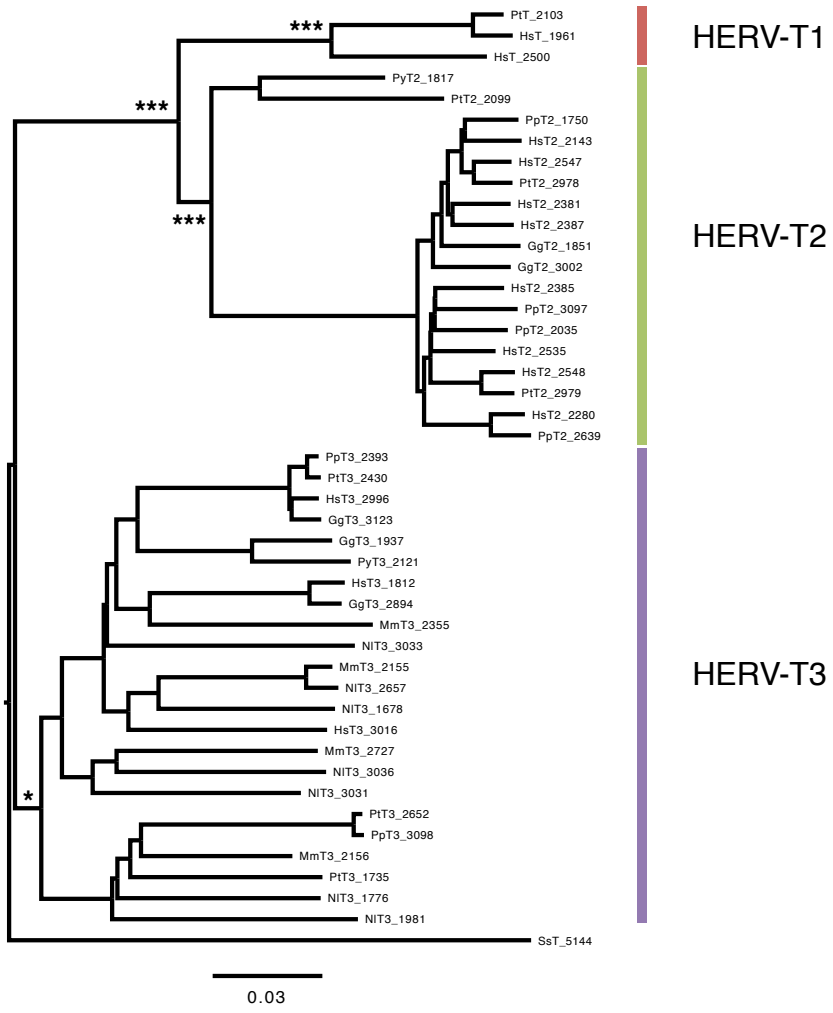


Figure 3.1: HERV-T proviruses cluster into three monophyletic clades.

Maximum Likelihood phylogenetic tree of 44 HERV-T proviral loci with a structure of the form LTR-*gag*-*pol*-*env*-LTR. The three distinct monophyletic clades are indicated with colors. The phylogenetic tree was rooted based on an out-group sequence (SsT_5144). Bootstrap support on internal nodes is indicated by asterisks: (*) > 80, (***) > 99. (1000 bootstrap replicates).

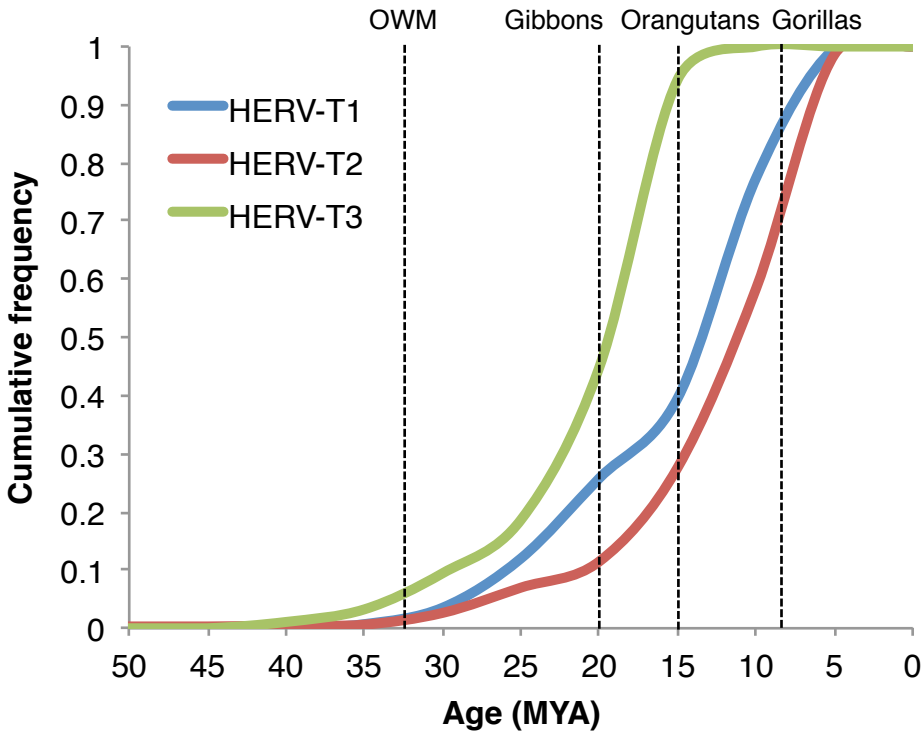


Figure 3.2: Dynamics of the HERV-T lineage through time.

Cumulative distribution of soloLTRs by its date of integration. Integration date approximated by adjusting the distance to their corresponding consensus sequence and a human neutral substitution rate of 2.2×10^{-9} substitutions per site per year (Lander et al., 2001) (see chapter II). Speciation dates for distinct primate groups or species are indicated by dotted lines. OWM: Old World Monkeys.

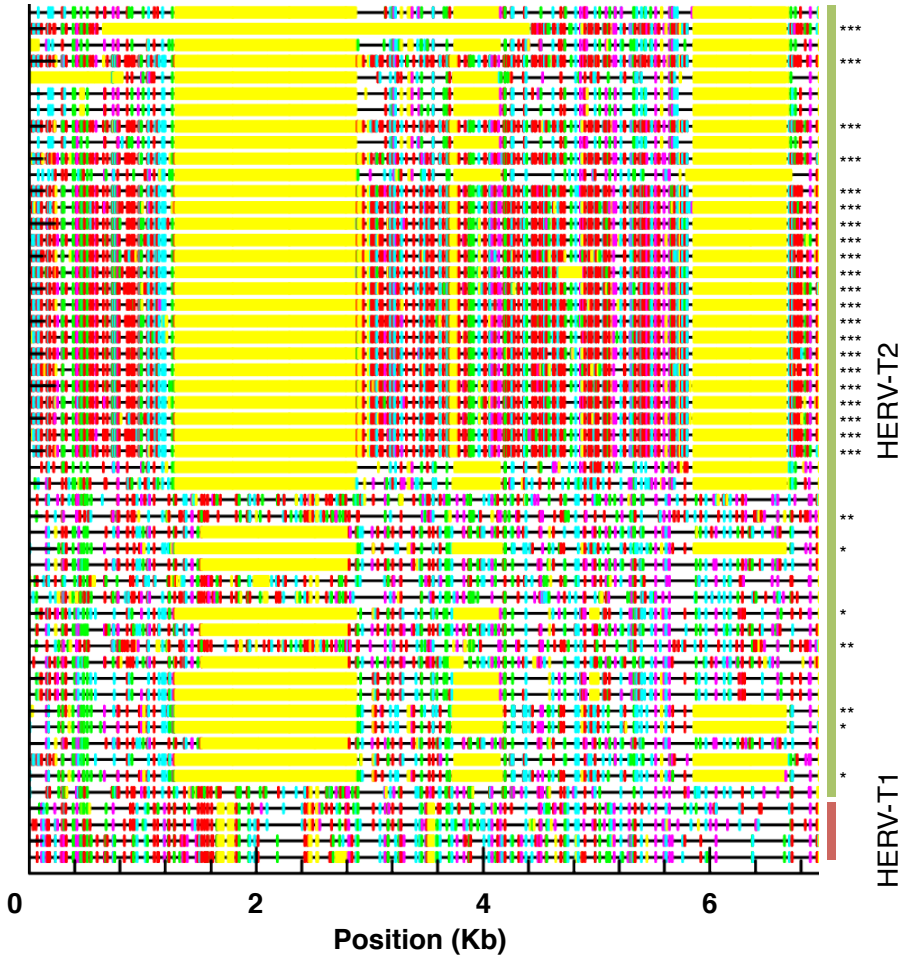


Figure 3.3: Hypermutation analysis of HERV-T1 and HERV-T2 proviral sequences.

Profile of G to A transitions of HERV-T2 and HERV-T1 proviral sequences containing fragments of coding genes (*gag-pol-env*). Proviral sequences were aligned to a consensus HERV-T sequence and its APOBEC3 mediated hypermutation profile was assessed using Hypermut 2.0 (Rose and Korber, 2000) with default parameters. Lines in red and cyan represent A3-derived G to A transitions (GG to AG and GA to AA respectively), whereas lines in green and magenta represent non APOBEC3-derived G to A transitions (GC to AC and GT to AT respectively). Lines in yellow indicate gaps compared to the consensus sequence. (*) p-value < 0.05, (**) p-value < 0.01, (***) p-value < 0.001

Analysis of a murine specific ERV lineage transcriptionally active in the mouse embryo.

As previously described in chapter I, MuERV-L is a highly abundant mouse specific endogenous retrovirus that continues to be transcriptionally active at an early stage of the mouse embryo (Kigami et al., 2003; Macfarlan et al., 2011; Macfarlan et al., 2012; Ribet et al., 2008b). Previous analyses have established that MuERV-L underwent two amplification bursts, one after the divergence of the Mus and Rattus genera around ~10 million years ago (MYA), and a more recent burst about 2 MYA (Costas, 2003). MuERV-L belongs to a large family of ERVs that has been active throughout the evolution of mammals (Benit et al., 1999; Cordonnier et al., 1995; Lee et al., 2013). RT phylogenetic trees show MuERV-L inside class III retrovirus, clustering together with spumaviruses and human ERV-L (HERV-L) and HERV-S. In contrast to their human counter part, most of MuERV-L copies have complete coding potential, encoding open reading frames (ORFs) for gag and pol (Benit et al., 1997). As other ERV-L elements, MuERV-L is characterized by the complete absence of an *env* gene that, coupled with its early transcription profile, suggest an entirely intracellular expansion through retrotransposon-like replication (Magiorkinis et al., 2012). In fact it has been shown that MuERV-L transcripts are able to form intracellular viral-like particles that accumulate in the ER (Ribet et al., 2008b).

In order to further investigate the evolution of this ERV lineage, we first aimed to describe the diversity of MuERV-L related sequences in the mouse genome

using DIGS. Currently there are two available complete genome assemblies: the Mouse Genome Reference Consortium build 38 (GRCm38 also known as mm10) corresponding to the C57BL/6J strain (Mouse Genome Sequencing et al., 2002), and the whole genome shotgun (WGS) assembly from Celera that corresponds to a mixture of 5 strains (129X1/SvJ, 129S1/SvImJ, DBA/2J, A/J and C57BL/6J) (Mural et al., 2002). We mined both genome assemblies by DIGS using separate Gag, Pol and LTR probes from the MuERV-L reference sequence (GenBank: Y12713) (Benit et al., 1997), followed by consolidation of the resulting hits.

Table 3.3: MuERV-L sequences identified in different mouse genome assemblies

Assembly	Strain	Reference	Total	solo LTRs	Provirus	Fragments
GRCm38	C57BL/6J	Mouse Genome Sequencing et al., 2002	2971	1588	719	664
Mm_Celera	Mixture	Mural et al., 2002	2768	1775	220	773

Overall we found less than 3,000 MuERV-L elements in the mouse genome involving three main types of structures (Table 3.3 and Figure 3.4). Proviruses flanked by two LTRs show the biggest discrepancy between genome assemblies finding 220 proviruses in the Celera assembly and 719 in GRCm38 (Table 3.3). This observation might be explained by the different mouse strains and/or

assembly methods used for each project. As expected, soloLTRs account for more than 50% of the elements present in the mouse genome (Table 3.3). The remaining MuERV-L structures, those composed of internal sequences with or without an associated LTR, roughly represent ~25% of all elements (Table 3.3). Given that GRCm38 is well supported by external and internal annotations we utilized this assembly to analyze the distribution of MuERV-L elements in the genome (Figures 3.4 and 3.5). We found an even distribution of MuERV-L elements in mouse chromosomes with a slightly higher density of soloLTRs in the X chromosome (0.77 soloLTRs/Mb) and of provirus in chromosomes 7, 12 and 13 (~0.35 Proviruses/Mb) (Figure 3.4A). The large heterochromatin region in chromosome Y might explain the relative low density of MuERV-L elements in this chromosome, possibly by preventing new integrations in this region (Figure 3.4A). The vast majority of elements were found in intergenic regions while ~7% of elements were found inside introns (Figure 3.4B). The remaining elements (0.94%) were found in non-coding RNA genes; un-translated regions and only 4 LTRs were found overlapping coding exons (corresponding to *pramel5*, *d5ertd579e*, *rpia* and *naalad2* genes) (Figure 3.4B). Such distribution differs significantly compared to randomized controls (p-value < 0.001). Similar scenarios are showed by soloLTR, proviral and fragment elements separately.

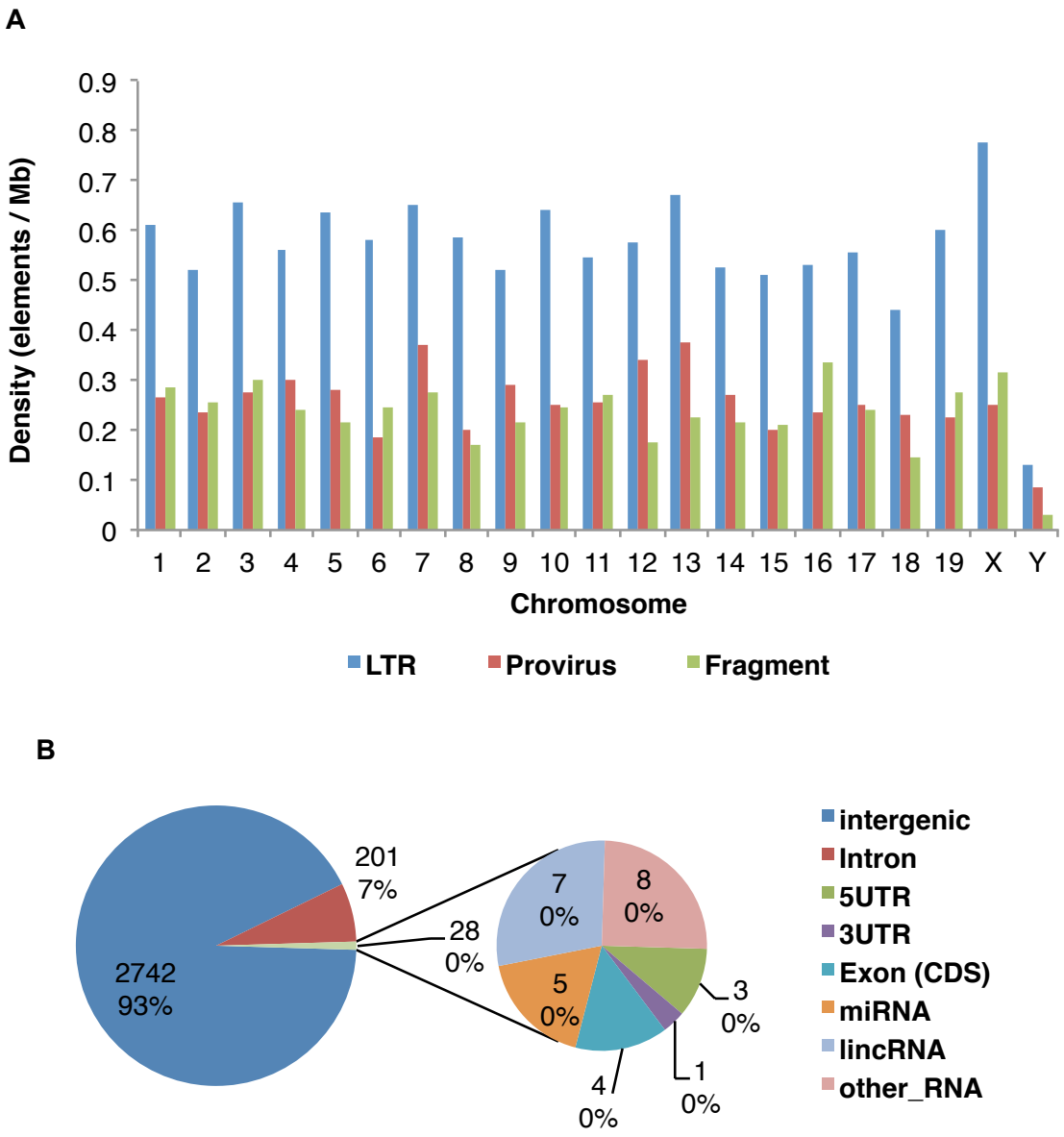


Figure 3.4: Distribution of MuERV-L elements in the mouse genome.

(A) Density of distinct MuERV-L structures in mouse chromosomes. Density measured as the number of elements divided by the size of the chromosome. **(B)** Distribution of MuERV-L structures in distinct mouse genomic features. The number of elements in each feature as well as the corresponding percentages is indicated. Data from a DIGS screen on GRCm38/mm10 (Mouse Genome Sequencing et al., 2002).

We noticed that the distribution of soloLTRs integrated in genes or intergenic regions also varies significantly from randomized controls (p-value < 0.001) (Figure 3.5). By classifying each soloLTR by integration date (see chapter II), we observed a negative trend between the age of the integration and its presence in genes (Figure 3.5). This difference is statistically significant when comparing soloLTRs of 0-2 MY old and those of 6-10 MY old (p-value = 0.01243) (Figure 3.5). Proviral sequences also show a negative trend in genes between elements involved in the first or second expansion (Figure 3.5). Although the distribution of these proviral elements is not significantly different, they differ significantly from randomized controls (p-values < 0.01) (Figure 3.5). These observations suggest a selective pressure against fixation of MuERV-L elements in genic regions. Moreover this selective pressure seems to be more relaxed for younger integrations. Further paleovirological analyses were performed on MuERV-L and are described in the following chapter.

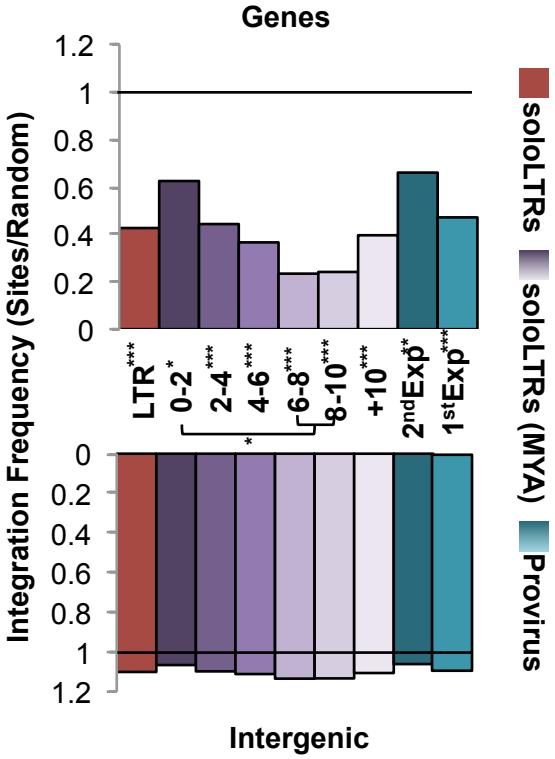


Figure 3.5: Distribution of MuERV-L elements in genic or intergenic regions in the mouse genome relative to random controls.

The measured value indicates the percentage of MuERV-L elements in each population divided by that of the random controls (10,000 random controls of 500nt for soloLTRs comparisons and 5,000 random controls of 6,500nt for proviral comparisons). The horizontal black line indicates no difference between the ratio of MuERV-L elements in each population and that of the controls. (*) p-value < 0.05. (**) p-value < 0.01. (***) p-value < 0.001. P-values are based on chi-squared goodness-of-fit or contingency table tests.

Summary

These analyses showed the capability of DIGS to identify complex proviral structures despite the numerous mutational processes by which these sequences have evolved. Consolidation of proviral loci also revealed that most ERV lineages analyzed have a significant amount of mosaic species. The mosaicism observed could account for true recombination events, particularly between closely related lineages, but also between distant ones. Overall this screening tool, together with previously published programs and functions, are instrumental in guiding us through the process of identifying suitable candidate sequences in which to perform paleovirological analyses to approximate a possibly infectious ancient retroviral sequence.

Chapter IV. Ancestral reconstruction of an infectious murine ERV

Given the characteristics of MuERV-L we decided to construct an infectious sequence using the information gained from the screening of MuERV-L related sequences in the mouse genome described in the previous chapter.

Strategy for ancestral reconstruction of MuERV-L

As mentioned earlier, a previous study (Costas, 2003) had determined that MuERV-L elements involved in the most recent and prolific expansion (~2 MYA) differentiate from the ones involved in a previous expansion (~10 MYA) by a 33 nucleotide in-frame deletion in the 5' half of the *gag* ORF (Figure 4.1A). In order to infer the sequence of the most recent infectious ancestor, while still accounting for informative variation, we decided to approximate the putative infectious ancestral sequence of MuERV-L by reconstructing the *gag* sequence separated from the rest of the genome (LTR and *pol*). For the reconstruction of the LTRs and the *pol* ORF we selected 95 complete (LTR-*gag*-*pol*-LTR) proviruses closely related to the reference MuERV-L sequence (Benit et al., 1997) to guide a maximum likelihood (ML) reconstruction of the root node (*pol* n96) and an identical pair of LTRs (Figure 4.1B and 4.1D).

To select *gag* sequences only belonging to the most recent expansion we downloaded 230 *gag-pol* containing loci (identified by DIGS) that were present in both Celera and GRCm38 assemblies. After aligning those loci to the reference

sequence we identified a monophyletic clade of 40 elements that exclusively contained a 33-nucleotide in-frame deletion in the 5' half of their *gag* ORFs. The loci in this clade were selected to guide a maximum likelihood (ML) reconstruction of the root node (*gag* n377) (Figure 4.1C and 4.1D). The combined LTR, *gag* and *pol* ancestral sequences were corrected for possible mutations derived by deamination of methylated CpG dinucleotides to create a ~2 MY ancestral MuERV-L sequence (ancML) (Figure 4.1D and Figure 4.2) (see chapter II).

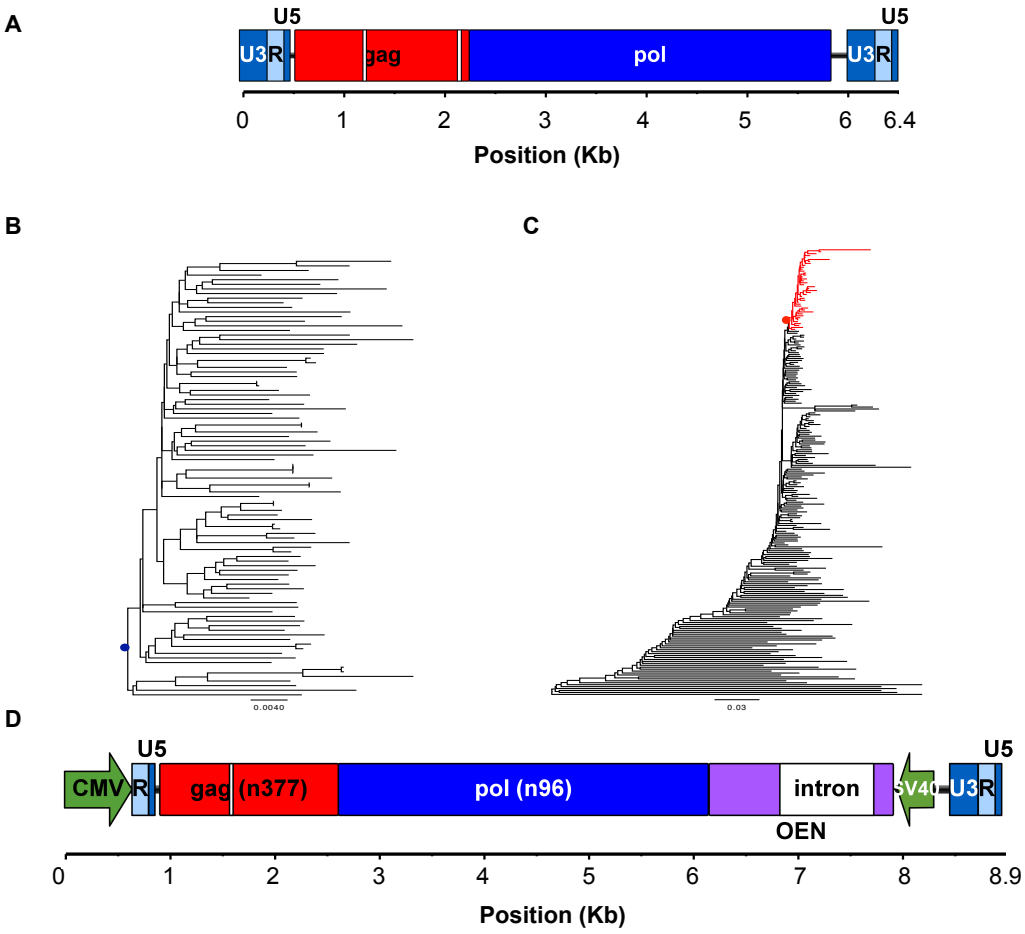


Figure 4.1: Strategy and reconstruction of an ancestral MuERV-L genome.

(A) Schematic representation of the structure of MuERV-L elements. **(B)** Maximum likelihood phylogenetic tree of 95 LTR-*gag*-*pol*-LTR MuERV-L elements in the mouse genome. Blue circle denotes the ancestral node reconstructed by baseml (*pol* and LTR, node 96). **(C)** Maximum likelihood phylogenetic tree of 230 gag-pol containing MuERV-L elements in the mouse genome. The monophyletic red clade contains only elements with a 33-nucleotide deletion in gag at position 671 with or without the additional deletion in gag at position 1597. Red circle denotes the ancestral node reconstructed by baseml (gag, node 377). **(D)** Organization of the ancML construct. Green arrows indicate promoter sequences. OEN: *neo* gene in reverse orientation from the ancML transcription. White boxes represent the 33 and 39 nucleotides deletions in gag at positions 671 and 1597, respectively.

Figure 4.2: Nucleotide sequence and translation products of ancML.

LTR sequences are shown in bold italics. Nucleotide and protein sequence of *gag* and *pol* are indicated in red and blue, respectively (amino acid single letter code, (*) represents stop codons). The 33-nucleotide deletion in *gag* is shown with a magenta triangle. The position of the 39-nucleotide deletion in *gag* is highlighted in magenta. The RT active site is highlighted in yellow. The PBS is indicated in violet, the polypurine tract in red, the TATA box in green and the polyadenylation site in bright blue. Protein positions are indicated in bold, while nucleotide positions are not.

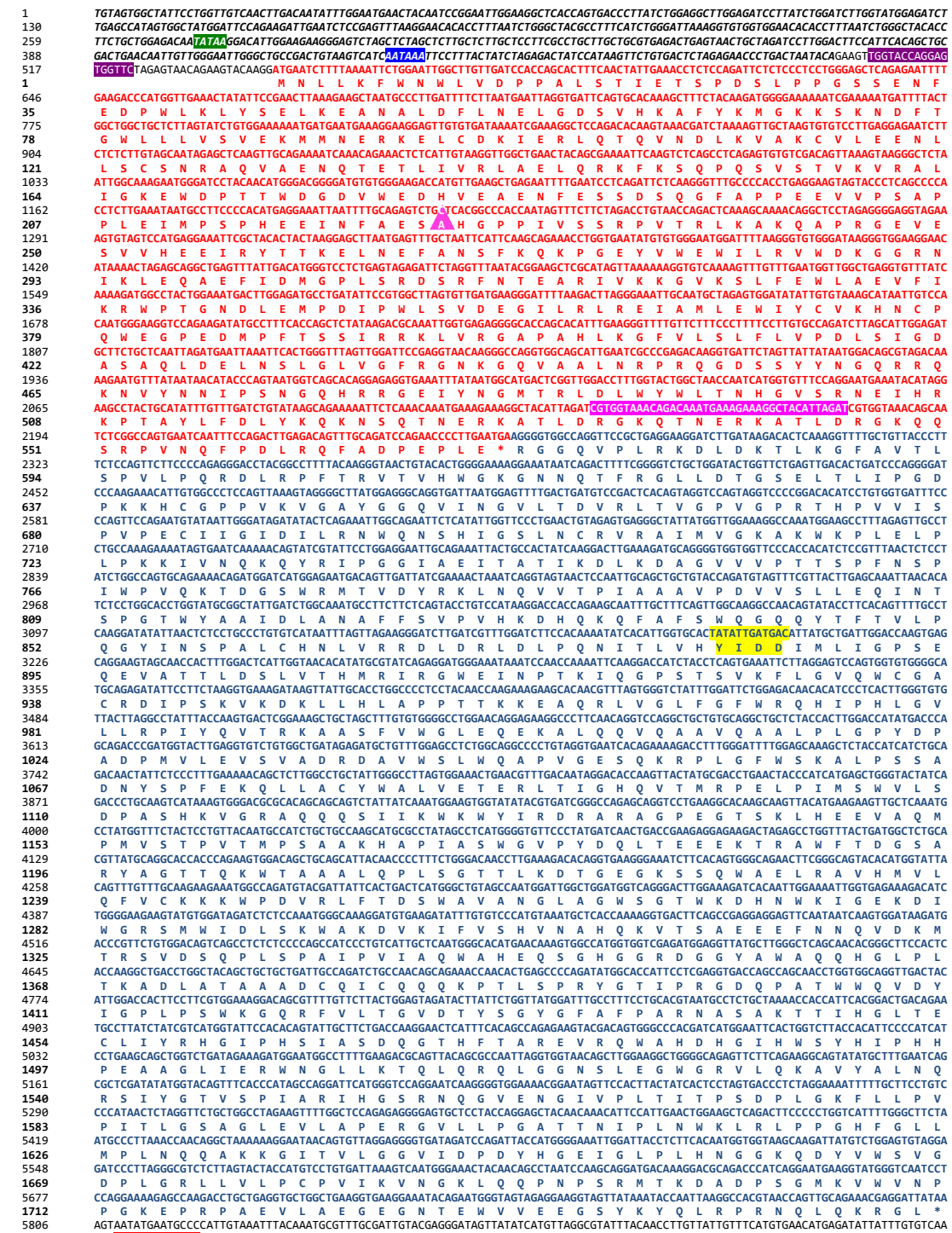


Figure 4.2: Nucleotide sequence and translation products of ancML.

ancML is infectious and its activity is dependent on a functional reverse transcriptase.

In order to assess the infectiousness of the reconstructed ancestral sequence, we cloned the full-length ancML sequence into an expression vector by replacing the U3 region of the 5' UTR by a CMV promoter. A G418 resistant (*neo*) gene controlled by a separate SV40 promoter and interrupted by an intron was further inserted at the end of pol and before the 3' LTR promoter (Figure 4.1D). The splice donor and acceptor sites of the intron are in the same orientation as the ancML transcription, in a way that only if ancML undergoes reverse transcription and integrates in the host cell genome, the intron is removed and a functional G418 resistance protein is expressed (Moran et al., 1996).

We then determined if such a plasmid could produce G418 resistant colonies in cell lines. For this purpose we tested a set of 16 cell lines: 8 from primates, 6 from rodents, 1 from a carnivore and 1 from an avian species (Table 4.1). Despite efficient transfection in the majority of the cells tested, the ancML expression plasmid was able to derive G418 resistant colonies only in Chinese hamster ovary K1 (CHO) cells and the related pgsA cells, and to a much lower extent in Vero cells (Figure 4.3A and Table 4.1). Surprisingly a control plasmid expressing a LINE1 element under the same replication dependent *neo* gene (L1.3 plasmid) (Moran et al., 1999) failed to produce G418 resistant colonies in 7 of the cells tested, including Vero (Table 4.1).

To further investigate the nature of these G418 resistant colonies, we mutated the ancML RT active site from YIDD to AIAA (Figure 4.2). This RT mutation completely abolished the production of G148 resistant colonies in CHO cells (Figure 4.3B). Additionally we observed that the ancML and L1.3 constructs are slightly sensitive to the presence of azidothymidine (AZT), a retroviral RT inhibitor (Figure 4.3B), indicated by the reduction of G418 resistant colonies. However, increasing concentrations of AZT also show a modest toxicity in CHO cells transfected with a control plasmid expressing the *neo* gene, particularly at a concentration of 500uM (Figure 4.3B). This observation may obscure the real effect of this inhibitor on ancML and L1.3 replication.

As we mentioned previously, the fate of the intron that interrupts the *neo* gene is in direct correlation with the appearance of G418 resistance colonies. We isolated genomic DNA (gDNA) from CHO cells transfected with the ancML plasmid or an empty vector and determined the fate of the intron by PCR (Figure 4.3D). In CHO cells transfected with the ancML plasmid the vast majority of the amplified sequences correspond to the proper processing of the intron interrupting the *neo* gene. We also observed the presence of molecules (in low abundance) whose size corresponds to the retention of the intron, possibly accounting for traces of the transfected plasmid (Figure 4.3D). It is important to notice that the primer-binding site (PBS) is separated from the 5' LTR U5 region by a five-nucleotide linker sequence (Figure 4.2 and Figure 4.4B). This separation is uncommon between exogenous retroviruses and in fact is only

observed in one other ERV, HERV-E (Repaske et al., 1985). Elimination of this five nucleotides in the ancML sequence resulted in a ~4 fold reduction in the number of G418 resistant colonies and the incorrect processing of the intron interrupting the *neo* reporter gene (Figure 4.3C), suggesting a role in MuERV-L replication. Overall, these results indicate that the reconstructed ancestral MuERV-L sequence is infectious and is able to undergo reverse transcription and integration upon transfection into CHO cells.

Table 4.1: Cell lines tested for replication of ancML.

Cell Line	Organism	Order	GFP[#]	NEO	L1.3	ancML
DF-1	*Gallus gallus*	Galliforme	***	+++	++	-
CRFK	*Felis catus*	Carnivora	**	+++	+	-
CV-1	*Cercopithecus aethiops*	Primate	*	++	+	-
Vero	*Cercopithecus aethiops*	Primate	***	+++	-	+
HT1080	*Homo sapiens*	Primate	**	++	+	-
HOS	*Homo sapiens*	Primate	*	++	-	-
Huh7.5	*Homo sapiens*	Primate	*	++	-	-
TE671	*Homo sapiens*	Primate	-	-	-	-
MRC-5	*Homo sapiens*	Primate	-	-	-	-
HeLa	*Homo sapiens*	Primate	***	+++	++	-
pgsA745	*Cricetulus griseus*	Rodent	**	++	+	++
CHO-K1	*Cricetulus griseus*	Rodent	**	++	+	++
MusDunni	*Mus dunni*	Rodent	***	+++	+	-
SC-1	*Mus musculus*	Rodent	**	++	-	-
NIH3T3	*Mus musculus*	Rodent	**	++	++	-
Rat2	*Rattus norvegicus*	Rodent	-	-	-	-

All cell lines were transfected using lipofectamine 2000.

#A plasmid expressing GFP was utilized as a transfection control.

Percentage of GFP positive cells: (-) 0%, (*) < 10%, (**) 10-50%, (***) > 50%.

Number of G418 resistant colonies: (-) None, (+) ≤ 10, (++) > 10, (+++) > 50.

Figure 4.3: Infectivity of ancML is dependent on reverse transcription.

(A) G418 resistant colonies on 15 cm cell culture plates derived from CHO cells transfected with plasmids expressing ancML, L1.3 or an empty vector. **(B-C)** G418 resistant colonies of CHO cells cultured with increasing amounts of AZT. CHO cells were transfected with plasmids expressing a *neo* gene (NEO), L1.3, ancML, (B) an ancML construct with inactivating mutations in the RT active site (ancML-RTmut), or (C) an ancML construct with the 5nt linker sequence deleted (ancML ΔGAAGT). AZT treatment in (B) lasted for 2 days before G418 selection. Data from 3 independent experiments. **(D)** PCR amplification of the *neo* gene in genomic DNA (gDNA) of CHO cells transfected with a plasmid expressing ancML, ancML ΔGAAGT or an empty vector. A scheme of the PCR amplification strategy is shown on top. The usage of DNA from plasmid or from CHO gDNA as well as a water control is indicated. L: molecular weight ladder. In all experiments CHO cells were transfected with 1μg of plasmid DNA and placed on G418 selection 3 dpt. Cells were cultured under selection for an additional 10 days before treatment with paraformaldehyde (PFA) and crystal violet staining (A, B and C), or gDNA extraction (D).

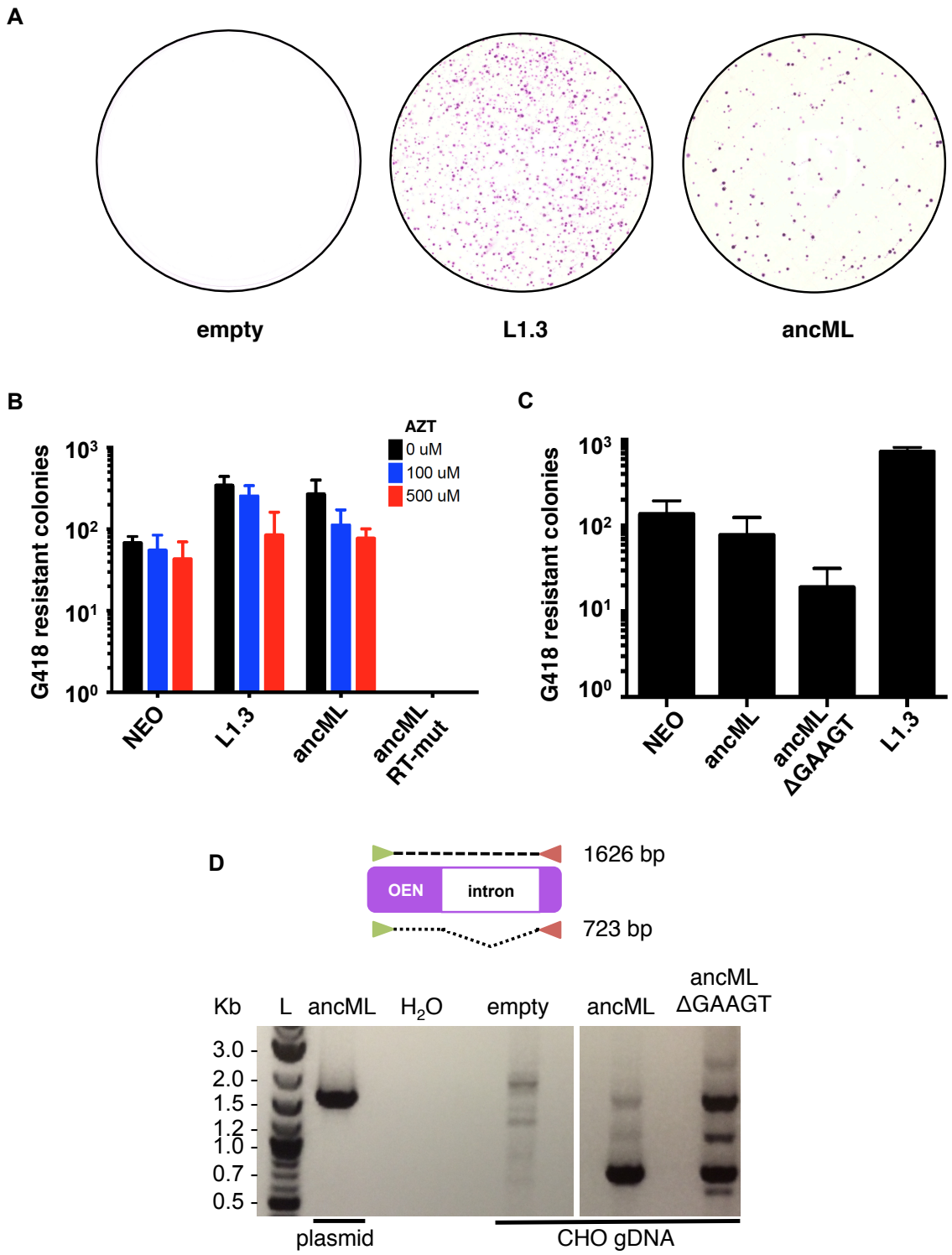


Figure 4.3: Infectivity of ancML is dependent on successful reverse transcription.

Analysis of ancML integration sites in CHO cells.

We next aimed to determine the pattern of integration of ancML when transfected into CHO cells. For this purpose we utilized a genome walker technique (Clontech) to PCR amplify integration sites using primers specific to the flanking LTRs and an adaptor sequence (Figure 4.4A). We extracted gDNA from G418 resistant single cell clones of CHO cells transfected with ancML. The gDNA was further digested with EcoRV (absent from ancML) and ligated to adaptors. Nested PCR reactions on gDNA from these cells showed bands of different sizes and intensities consistent with multiple integration sites per cell (Figure 4.4B). We selected the most intense bands and cloned them into a TOPO vector (Life technologies) for sequencing. All of the sequences amplified from the 5' flanking side corresponded to neo-intron ancML sequences from priming on the homologous U3 sequence on the 3' LTR (Figure 4.4A). Similarly, the majority of the sequences amplified from the 3' flanking side corresponded to plasmid DNA and leader-gag ancML sequences from priming on the homologous R sequence on the 5' LTR (Figure 4.4A). Although there is no EcoRV site in the ancML sequence, possible shearing of gDNA during processing of the samples might have resulted in the amplification of sequences derived from the opposite LTR (Figure 4.4A). Despite this fact, we were able to PCR amplify, clone and sequence 26 bona-fide 3' flanking integration sites (Figure 4.4C and Table 4.2).

Initial inspection revealed that 10 of the 26 integration sites included a portion of the leader sequence containing different lengths of the PBS and the five-

nucleotide linker sequence (Figure 4.4C). During reverse transcription and after the synthesis of the plus-strand strong-stop DNA (+sssDNA), RNase H removes the primer tRNA, thereby exposing sequences on the +sssDNA that are complementary to the PBS which will guide the second strand transfer (Champoux and Schultz, 2009; Coffin et al., 1997) (Figure 1.5). Inefficient removal of the tRNA primer might result in the synthesis of additional PBS sequences after the 3' LTR on plus strand viral DNA, which might explain the integration pattern we observe for ancML.

The CHO genome (CriGri_1.0) is currently assembled to the scaffold level and has been annotated by distinct *de novo*, expression-based and homology gene prediction systems (Xu et al., 2011). In order to determine if ancML integration events had a preference for particular regions in the genome we mapped the gDNA of the 26 integration sites in the Chinese hamster genome using the UCSC genome browser (Figure 4.4D and Table 4.2) (Kent et al., 2002; Xu et al., 2011). Similarly to the MuERV-L loci present in the mouse genome, the majority of the ancML integration sites (19 sites) corresponded to intergenic regions, 5 sites corresponded to introns and one to an exon (exon 3 of *znf462*) (Table 4.2). The remaining site could not be classified as intergenic or in genes because it mapped to multiple scaffolds. Regarding repetitive sequences of the 26 integration sites 10 integrated in elements corresponding to SINE (4), LINE (4) and ERV-L (2) elements (Table 4.2). In order to account for biases due to location and density of EcoRV restriction sites, we compared the distribution of

the 26 ancML integration sites to matched random controls. Each integration site was matched to three random genomic locations that were at the same distance to an EcoRV site as the one showed in the flanking CHO DNA sequence (Marshall et al., 2007). We observed that the distribution of the integration sites within genes or in intergenic regions, as well as for the integration in/outside repetitive sequences, did not differ significantly to randomized controls (p-value = 0.97 and 0.56 respectively) (Figure 4.4D). Although the distribution of the sequenced ancML integration sites and the distribution of MuERV-L elements in the mouse genome are different (Figures 3.5 and 4.4D), statistical significance could not be established due to the reduced sample size. These results suggest that the proportions of ancML integration sites we observed for different genomic regions correspond closely to what we would expect if the integration events would occur by chance, arguing against a particular preference for ancML integration.

Figure 4.4: Integration site analysis of ancML in CHO cells.

(A) Schematic representation of the genome walker strategy used to sequence ancML integration sites. Forward and reverse primers (green and red triangles) were designed to anneal to the R and U3 region of the LTRs, respectively. Additional primers (white triangles) were designed to anneal to an adaptor sequence (white box). True EcoRV restriction sites are depicted by an orange thunder. An accidental break in the middle of the provirus is depicted by a purple thunder. Additional adaptors might be ligated to accidental breaks and primer anneal might result in the amplification of internal parts of the provirus (dotted lines). CHO gDNA is represented by a grey line. Red and green lines indicate 5' and 3' integration sites, respectively. **(B)** Example of a genome walker experiment to determine the 3' flanking sequence of ancML integration events on 15 single cell clones resistant to G418. Nested PCRs were performed on gDNA of single cell clones of CHO cells transfected with a plasmid expressing ancML and previously digested with EcoRV and ligated to adaptors. Forward primers were designed to anneal to the R region of the 3'LTR and the reverse primers to an adaptor sequence. L: molecular weight ladder. **(C)** Sequences of 26 ancML integration sites in the CHO genome. Sequences of the Chinese hamster Leucine (TAA) tRNA as well as the ancML U5-PBS region are included in the first and second rows, respectively. Sequence from the U5 region of the ancML LTR is indicated in bright blue. The 5nt linker sequence is indicated in black. The PBS sequence is indicated in violet. CHO genomic sequence is indicated in bold. Residues highlighted in green indicate conservation in 30% of all sequences. **(D)** Distribution of ancML integration sites in genic, intergenic, or repeat regions relative to matched random controls. The measured value indicates the percentage of ancML integration sites in each population divided by that of the matched random controls (each integration site was matched to three random genomic sequences equidistant to the EcoRV site where the adaptor was ligated). The horizontal black line indicates no difference between the ratio of ancML integration sites in each population and that of the matched controls.

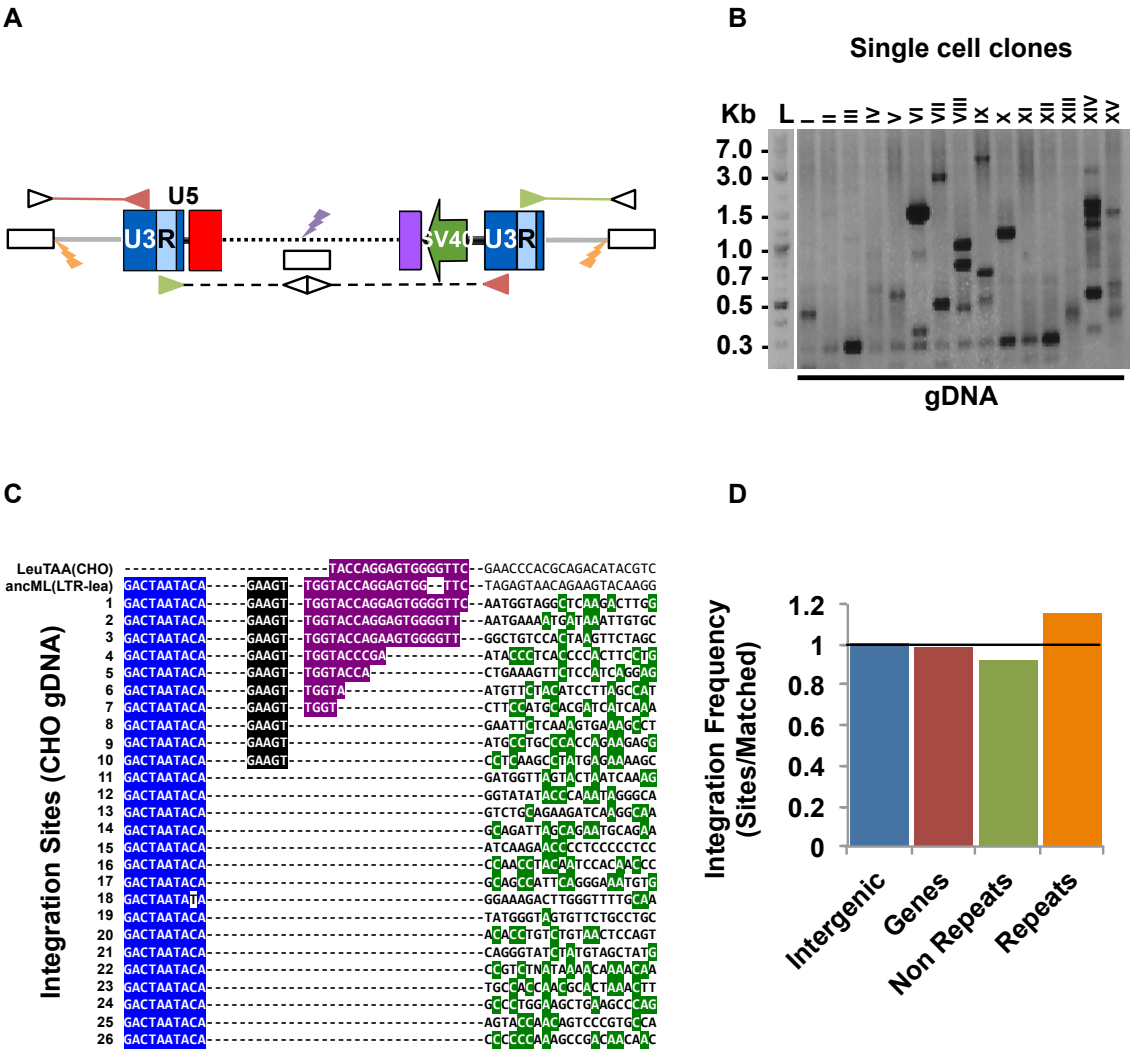


Figure 4.4: Integration site analysis of ancML in CHO cells.

Table 4.2: ancML integration sites on the CHO genome.

ID	Chr	Strand	Start	End	Region	Gene	Repeat	Notes
ancML1	KE381885	+	340717	340867	Intron	*JP064393*		
ancML2	KE379930	+	342872	343020	Intergenic			
ancML3	KE377538	+	145273	145423	Intergenic		L1 L1M4b	
ancML4	KE378078	-	661954	662104	Intergenic		ERV-L	
ancML5	KE379441	-	45852	46002	Intergenic		ERV-L	
ancML6	KE380245	+	2923	3073	Intergenic		L1 L1MA4	
ancML7	KE383346	-	192119	192269	Intergenic			
ancML8	KE381812	-	1057294	1057444	Exon	*ZNF462*		
ancML9	KE382795	-	1876320	1876468	Intergenic			
ancML10	KE382060	+	6750448	6750563	Intergenic		SINE B4	
ancML11	KE379127	+	588151	588301	Intron	*JP048133*		
ancML12	NA	NA	NA	NA	NA		L1 L1MA8	Multiple chromosomal locations
ancML13	KE382159	+	1085607	1085757	Intergenic		L1 Lx8	
ancML14	KE379529	+	298519	298627	Intergenic			
ancML15	KE376373	-	986092	986242	Intergenic			
ancML16	KE379280	-	800088	800238	Intergenic			
ancML17	KE380852	-	550987	551137	Intergenic			
ancML18	KE381723	+	2004500	2004650	Intergenic			
ancML19	KE377309	+	548867	549017	Intergenic			
ancML20	KE383226	+	3723754	3723904	Intron	*JI876162*	SINE B4a	
ancML21	KE382440	-	1741236	1741386	Intron	*JI874119*	Alu	
ancML22	KE378369	+	173291	173432	Intergenic		SINE B4	
ancML23	KE380364	-	88	238	Intergenic			
ancML24	KE379482	+	230888	231036	Intergenic			
ancML25	KE379368	+	220959	221109	Intron	*JI886141*		
ancML26	KE379377	-	178559	178709	Intergenic			

ancML is sensitive to host innate antiviral defenses

We then determined the potential for ancML to be inhibited by the mouse innate immune defenses. In order to address if ancML replication is affected by the interferon response, we transfected CHO cells with plasmids expressing ancML, L1.3 or a *neo* gene and cultured them with media containing increasing amounts of mouse IFN-α for two days before G418 selection (Figure 4.5A). We noticed a reduction in G418 resistant colonies in CHO cells expressing ancML or L1.3, but not a control *neo* gene. While the replication of L1.3 is reduced up to ~4 fold upon IFN-α treatment, there was a dose dependent effect on ancML reaching a ~20 fold reduction of G418 resistant colonies with 50u/ml of mouse INF-α (Figure 4.5A).

We also tested if specific innate immune effectors can have an effect on ancML replication. For this purpose we constructed CHO cells stably expressing mouse orthologs of known restriction factors *apobec3*, *mov10* and *samhd1* (isoform 2), that have been previously implicated in the control of endogenous and exogenous retroviruses and retro-elements (reviewed in (Rehwinkel, 2014)) (Figure 4.5B). Strikingly only mouse APOBEC3 (mA3) was able to restrict ancML showing a ~30 fold reduction in G418 resistant colonies compared to control CHO cells (Figure 4.5B). There was no effect of mouse MOV10 and SAMHD1 on ancML replication or of any of the restriction factors tested in L1.3 replication (Figure 4.5B). Surprisingly mouse MOV10 was not able to restrict the replication of L1.3, contrary to what has previously been observed for human MOV10 (Arjan-

Odedra et al., 2012). To further investigate the effect that mA3 might have on MuERV-L replication we determined the mutational profile of the 230 *gag-pol* containing MuERV-L elements that were used to derive ancML gag (Figure 4.1C) by using Hypermut 2.0 (Rose and Korber, 2000) (Figures 6C, 6D and 6E). For each MuERV-L element we calculated the ratio of G to A transitions in mA3-preferred motifs (5' G(AIG)(AIGIT) 3') and control sites (5' G(CIT)N 3' or 5' G(AIG)C 3') compared to a consensus sequence. Only three MuERV-L elements showed significant (p-value < 0.05) evidence of mA3 dependent hypermutation when no 5' context was enforced (Figure 4.5C). Because spontaneous deamination of methylated CpG dinucleotides can also produce G to A transitions, we excluded sites containing a C 5' to the mutated G. When these sites were excluded 10 MuERV-L elements showed a significant (p-value < 0.05) evidence for mA3 dependent hypermutation (Figure 4.5D). Only two MuERV-L elements were significant for mA3 dependent hypermutation in both analyses (p-value < 0.01) (Figure 4.5E). These results suggest that although ancML replication is strongly impaired by mA3 there is no sufficient evidence to conclude that mA3 dependent hypermutation had an inhibitory effect on MuERV-L during its evolution.

Figure 4.5: ancML is sensitive to mouse innate immune effectors.

(A) G418 resistant colonies of CHO cells cultured with increasing amounts of mouse IFN-α. CHO cells were transfected with plasmids expressing a *neo* gene (NEO), L1.3, or ancML and cultured with increasing amounts of mouse IFN-α for 2 days before G418 selection. Data from 3 independent experiments. **(B)** G418 resistant colonies of single cell clone CHO cells expressing mouse APOBEC3, MOV10 and SAMHD1 (isoform 2) genes. Single cell clone CHO cells were transfected with plasmids expressing a *neo* gene (NEO), L1.3, or ancML. Data from 3 independent single cell clone populations. **(C and D)** Mutational analysis of MuERV-L elements using Hypermut 2.0 (Rose and Korber, 2000). Ratio of G to A mutations from possible mA3 mutation sites (RD 3' to a G) versus control (YNIRC 3' to a G) with: (C) no 5' context or (D) excluding sites with a 5' C next to a G. 230 *gag-pol* containing MuERV-L elements in the mouse genome were compared to a consensus sequence. Data points in red and purple indicate MuERV-L sequences statistically significant hypermutated (p-value < 0.05). Data points in purple are statistically significant MuERV-L sequences shared between (C) and (D) (P-value < 0.01). **(E)** Profile of G to A transitions of two MuERV-L hypermutated proviral sequences compared to a consensus sequence. The profile of the reference MuERV-L sequence is shown for comparison (MLref, non significantly hypermutated). Lines in red and cyan represent mA3-derived G to A transitions (GG to AG and GA to AA respectively), whereas lines in green and magenta represent non mA3-derived G to A transitions (GC to AC and GT to AT respectively). Lines in yellow indicate gaps compared to the consensus sequence.

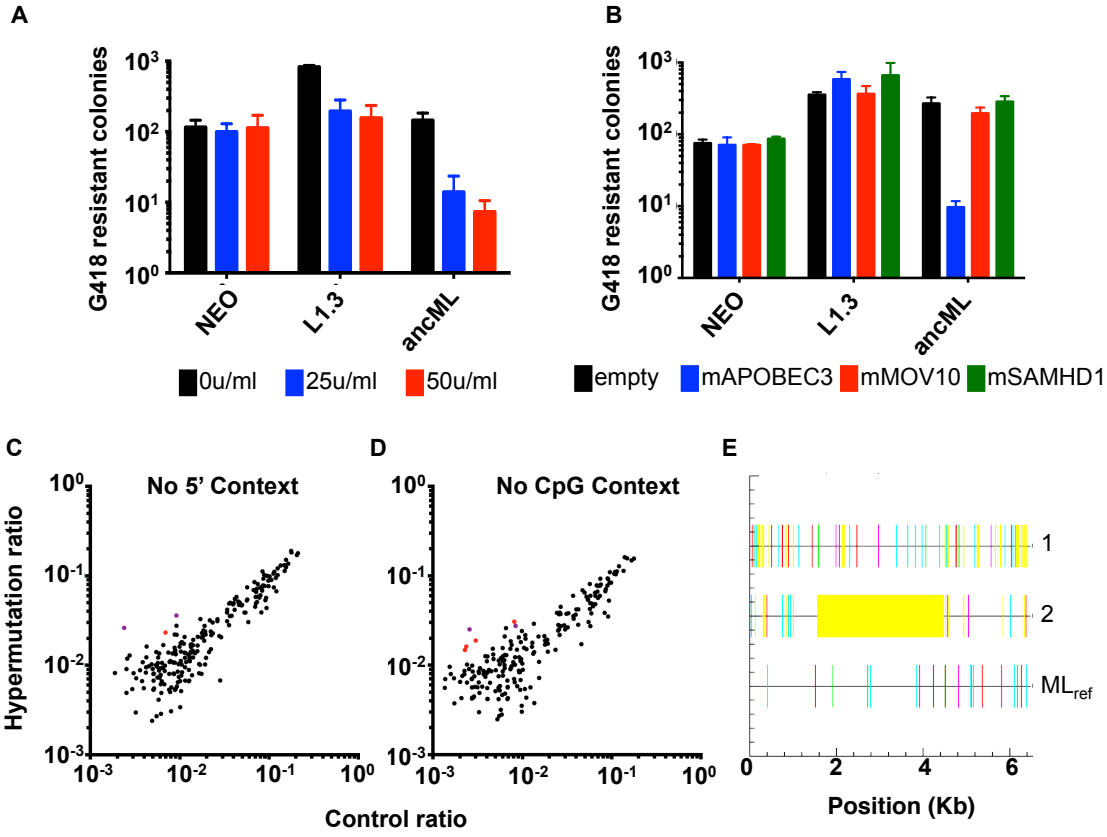


Figure 4.5: ancML is sensitive to mouse innate immune effectors.

Summary

In this chapter, we described the analysis and reconstruction of an infectious ancestral MuERV-L (ancML) sequence through paleovirological analyses of MuERV-L elements in the mouse genome. The resulting ancML sequence was infectious in CHO cells and its replication was dependent on efficient reverse transcription. We found that IFN-α could reduce ancML replication by ~20 fold. Additionally, we found that the expression of mouse APOBEC3 was able to restrict the replication of ancML. However, inspection of endogenous MuERV-L sequences suggested that the impact of APOBEC3-mediated hypermutation on MuERV-L evolution was limited. The development of a ~2 MYA ancestral MuERV-L sequence highlights the impact that retroviral elements have had in animal evolution, and the potential for the retroviral "fossil record" to uncover new biological processes.

Chapter V. Paleovirology of a human ERV

As mentioned in chapter III, HERV-T is a human ERV that is closely related to extant gammaretroviruses. The low copy numbers and the integrity of its *env* gene suggest for an expansion through reinfection mechanisms (Belshaw et al., 2005). These observations make HERV-T a suitable candidate to further investigate its evolution and to reconstruct a possible infectious ancestral sequence.

Ancestral reconstruction of HERV-T.

As described previously (chapter III), there are three classes of HERV-T sequences (Figure 3.1). Because most likely the HERV-T2 class originated from a subset of inactivated and hypermutated sequences that expanded through retrotransposon-like mechanisms (Figures 3.2 and 3.3), we discarded this class and center our analysis on the other two classes, HERV-T1 and the older HERV-T3. We downloaded proviral elements from these two classes that contained coding sequences flanked by two paired LTRs (LTR-*gag*-*pol*-*env*-LTR) in catarrhine primates. There were only 5 such elements belonging to the HERV-T1 class in the genomes of human and chimpanzee, whereas there were 27 HERV-T3 elements present in all catarrhine genomes analyzed. The proviral sequences were aligned to a consensus HERV-T reference sequence and were used to guide the construction of a ML phylogenetic tree (Figure 5.1). Analysis of the 5' and 3' flanking sequences revealed that there were only two and fourteen

unique HERV-T1 and HERV-T3 integration events (orthologous groups), respectively. The integration dates of these elements, approximated by the divergence of their paired LTRs, corroborated previous estimates for the expansion of HERV-T1 and HERV-T3 elements (Figures 3.2 and 5.1). Due to the paucity of HERV-T1 sequences we decided to use the proviral elements belonging to the HERV-T3 class to guide a ML ancestral reconstruction of its root node (Figure 5.1). The resulting ~32 MY old ancestral sequence was corrected for possible mutations derived from deamination of methylated CpG dinucleotides (see chapter II). The final ancestral sequence showed complete ORFs for all three genes (*gag*, *pol* and *env*), an unusually long leader sequence and an identical pair of LTRs.

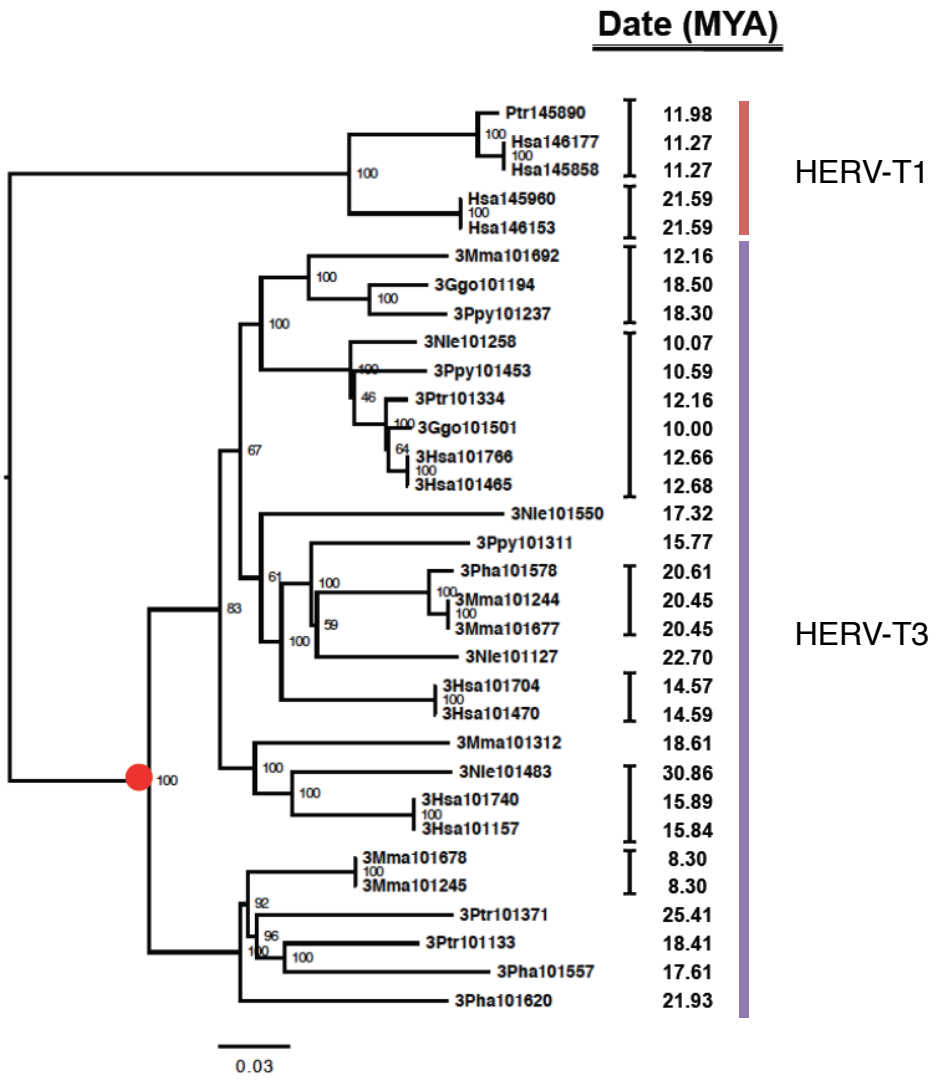


Figure 5.1: Phylogenetic tree of HERV-T1 and HERV-T3 proviral sequences in catarrhine primates.

LTR-*gag*-*pol*-*env*-LTR containing proviruses were aligned to a consensus HERV-T sequence and used to guide a ML phylogenetic tree. Black brackets represent orthologous groups. Dates of integration based on the divergence of their paired LTRs. Bootstrap support for internal nodes are indicated by numbers (10,000 bootstrap replicates). Red circle denotes the HERV-T3 ancestral node reconstructed by baseml.

The long leader sequence of HERV-T encodes a predicted transmembrane protein.

A closer look into the HERV-T ancestral and consensus sequence revealed a remarkably long leader sequence (1795nt), uncommon among gammaretroviruses and closely related ERVs (Figure 5.2). The retroviral un-translated leader sequence usually contains the PBS, regulatory RNA structures such as signals for packaging (Ψ), splicing, genome dimerization, and in some retroviruses an internal ribosome entry site (IRES) (Corbin and Darlix, 1996; Ohlmann et al., 2000). Additionally, an unconventional CUG codon present in the leader sequence of an exogenous gammaretrovirus, MLV, results in the production of an 88 amino acid extension of the canonical Gag protein, its translocation to the ER and its expression at the cell surface as a glycosylated transmembrane protein (MLV glycoGag) (Berlioz and Darlix, 1995; Pillemer et al., 1986; Prats et al., 1989). Moreover, the 815-nucleotide HERV-H leader sequence also encodes for an additional ORF upstream of *gag* with unknown function (Jern et al., 2005). In order to address if the unusually long HERV-T3 leader sequence could potentially encode for an additional ORF, we translated the ancestral leader sequence in all three frames and found one clear 726nt ORF 165nt upstream of the *gag* start codon. The preliminary translation of the "*pre-gag*" ORF resulted in a 242 amino acid protein with no sequence homology (nucleotide or protein) to any sequence present in public databases (Figure 5.3A). An ancestral sequence representing the HERV-T1 root node in Figure 5.1 also showed the

present of the *pre-gag* ORF but with a 132-nucleotide extension that encoded for a predicted tramsmembrane domain (Figure 5.3A). This extension was the result of a seven-nucleotide frameshift deletion in HERV-T1 proviruses compared to HERV-T3 (Figure 5.3A), which eliminated the stop codon. We then decided to revisit the HERV-T3 ancestral sequence to include those 43 codons that contained a transmembrane domain, and reassessed the state of particularly polymorphic sites, taking into account the variation present in HERV-T1 sequences (Figure 5.3A). The resulting 285-amino acid long pre-Gag protein (HTpG) was codon optimized for humans, synthesized and cloned into an expression vector that incorporated a HA-tag in its N-terminus (Figure 5.3A). Immunofluorescence assays of HA-tagged HTpG transfected into 293T cells shows a primarily intracellular and perinuclear localization in discrete structures most likely representing the ER and the membranous system (Figure 5.3B).

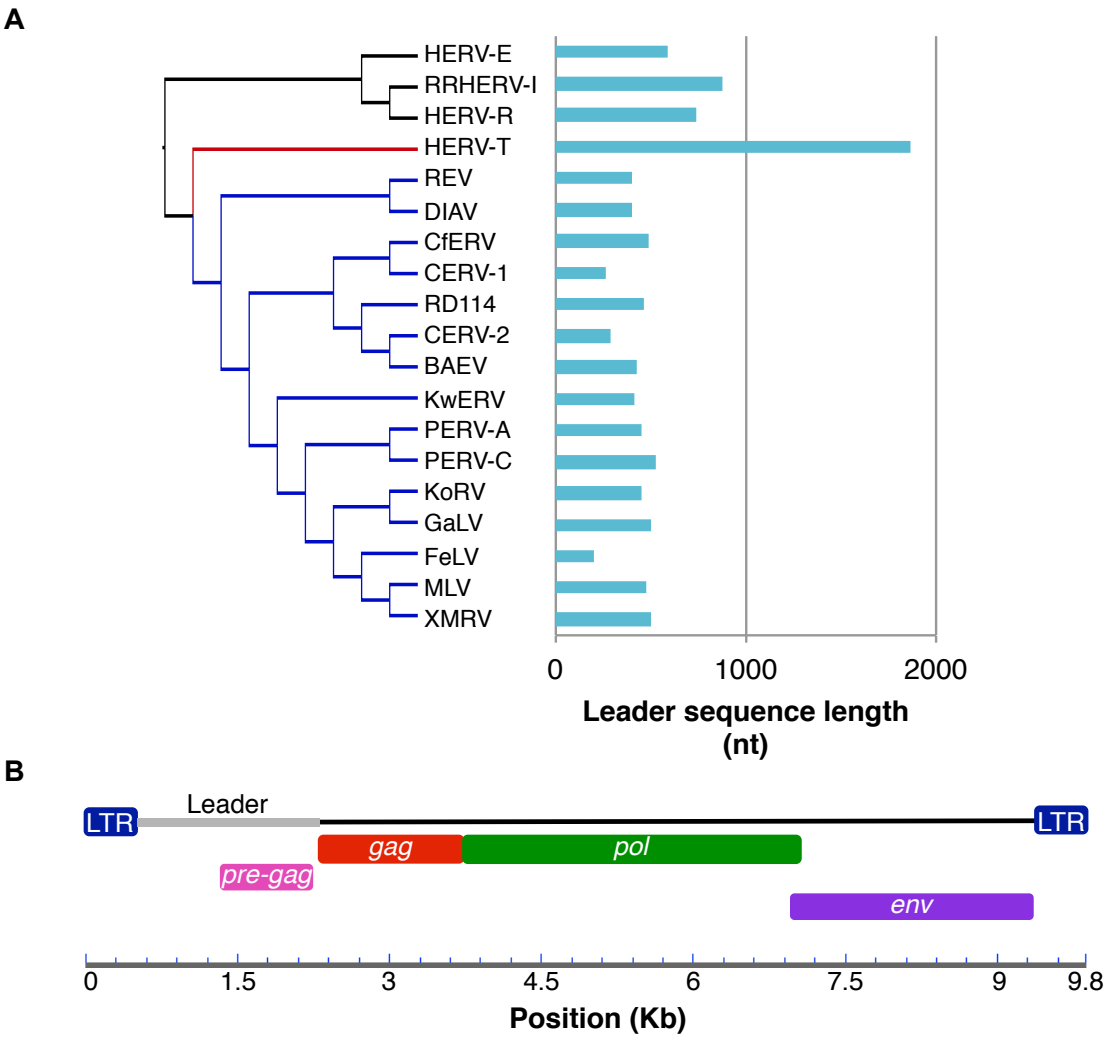


Figure 5.2: The leader sequence of the ancestral HERV-T3 genome contains a pre-gag gene.

(A) Length of the sequence between the end of the 5' LTR and the beginning of *gag* in nucleotides (nt), for selected endogenous and exogenous gammaretroviruses and closlely related ERVs. Phylogenetic relations are based on a distance (neighbor-joining) tree constructed from an alignment of consensus RT sequences. Gammaretroviruses have branches colored in blue. HERV-T branch is indicated in red. **(B)** Ancestral HERV-T3 genome. Coding sequences in different frames are indicated. Leader sequence is denoted in grey.

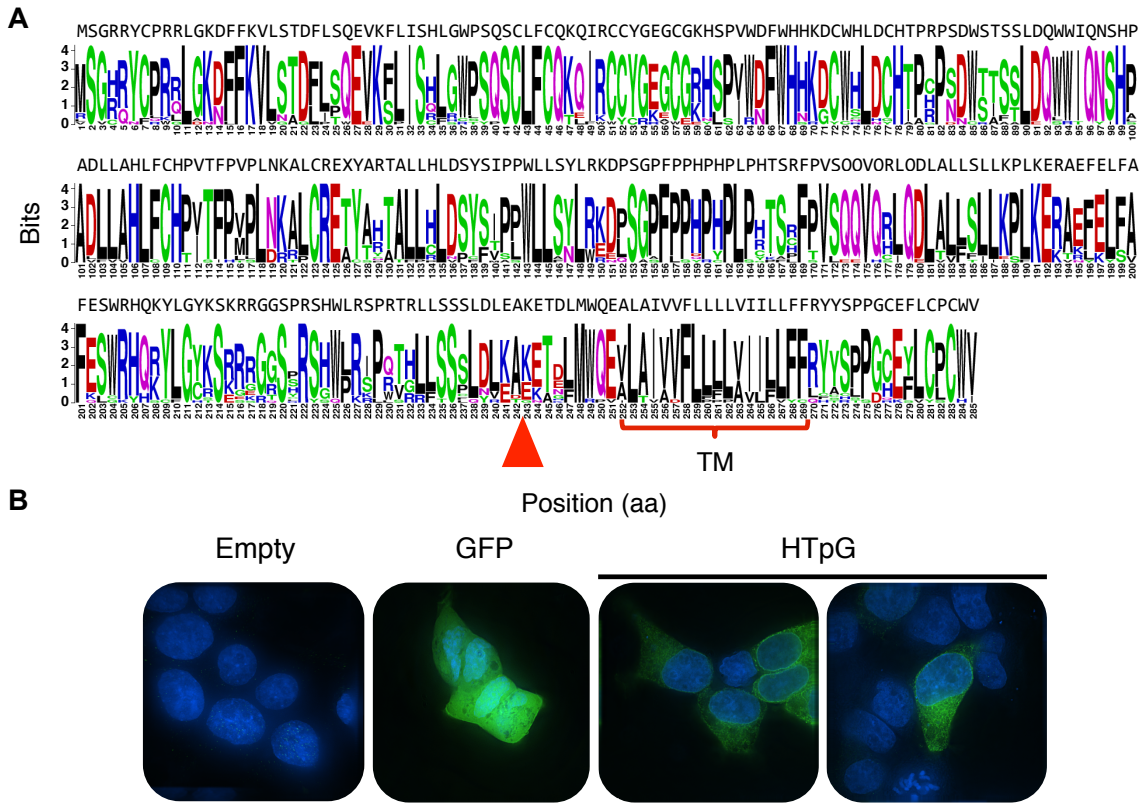


Figure 5.3: Sequence, expression and localization of the HERV-T pre-Gag protein.

(A) Logo plot of the translated *pre-gag* sequence in HERV-T3 proviruses. The total column height represents the sequence conservation on that site (in bits). The height of each residue represents its frequency at that particular site. The frameshift indel with respect to HERV-T1 proviruses is indicated by a red triangle. A red bracket indicates positions predicted to form a transmembrane domain. The HTpG final sequence is showed on top of the logo plots. **(B)** Subcellular localization of HTpG in 293T cells. Immunofluorescence of 293T cells transfected with and empty vector (Empty) or plasmids expressing GFP or an N-terminal HA-tagged HTpG protein. Fixed 293T cells were incubated with anti-HA antibodies. Nuclear DNA was labeled with DAPI.

The MLV glycoGag, mentioned before has been implicated to help viral infection and disease progression *in vivo* by a wide range of functions. MLV glycoGag enhances viral budding by redirecting virion assembly into lipid rafts (Low et al., 2007; Nitta et al., 2010) and inhibits the action of mouse APOBEC3 (Stavrou et al., 2013). More recently, its cytoplasmic tail has proven responsible for the rescue in infectivity of incoming Nef deficient HIV-1 particles (Pizzato, 2010; Usami et al., 2014) by inhibiting the action of the antiviral protein SERINC5 (Rosa et al., 2015; Usami et al., 2015). The action of SERINC5 and SERINC3 seems to be dependent on the envelope used (Pizzato, 2010; Rosa et al., 2015; Usami et al., 2015). In order to address if HTpG might replicate some of the functions associated with glycoGag we produced MLV particles pseudotyped with ecotropic or amphotopic MLV envelopes in 293T cells co-transfected with increasing amounts of a plasmid expressing HA-tagged HTpG or an empty vector (Figure 5.4). We were able to show a slight increase in infectivity of MLV particles pseudotyped with the amphotropic envelope (MLV-A), but not with the ecotropic envelope (MLV-E) (Figure 5.4A), in accordance with previous publications for MLV glycoGag (Pizzato, 2010). Although the effect of HTpG is small, the increase in infectivity of MLV-A pseudotyped particles is dependent on the amount of plasmid expressing HTpG (~2 fold increase in infectivity with the highest concentration of HTpG expressing plasmid) (Figure 5.4A). It is worth mentioning that there is a faint signal, about the size of the HA-tagged HTpG, in virion lysates produced from cells transfected with higher concentrations of the

plasmid expressing HA-tagged HTpG, perhaps suggesting minor incorporation of HTpG molecules into viral particles (Figure 5.4B).

Reconstruction of a functional ancestral HERV-T3 envelope.

As described in chapter III, HERV-T proviruses seem to have preserved the integrity of their *env* genes (Belshaw et al., 2005), which might be exploited in order to reconstruct a functional envelope. In order to approximate this possibly infectious *env* sequence, we refined the previously ~32 MY old predicted ancestral HERV-T3 *env* sequence by reassessing the state of particularly polymorphic sites taking into account the variation present in HERV-T1 sequences, in a way similar to the procedure used to reconstruct the HTpG protein (see also chapter II). The resulting 631-amino acid long gamma type-C refined ancestral envelope protein (ancHTenv) was codon optimized for humans, synthesized and cloned into an expression vector (Figure 5.5).

To determine if the reconstructed ancHTenv was infectious we produced MLV particles pseudotyped with this construct and tested its ability to infect a variety of animal cells (Figure 5.6A). Strikingly, ancHTenv was able to infect all primate, rodent and carnivore cell lines tested, with the exception of mouse NIH3T3, rat Rat2 and chicken DF-1 cells. The infectivity of the ancHTenv pseudoyped MLV particles was similar, or even better in some cell lines, than the infectivity of MLV particles pseudotyped with MLV-A (Figure 5.6A).

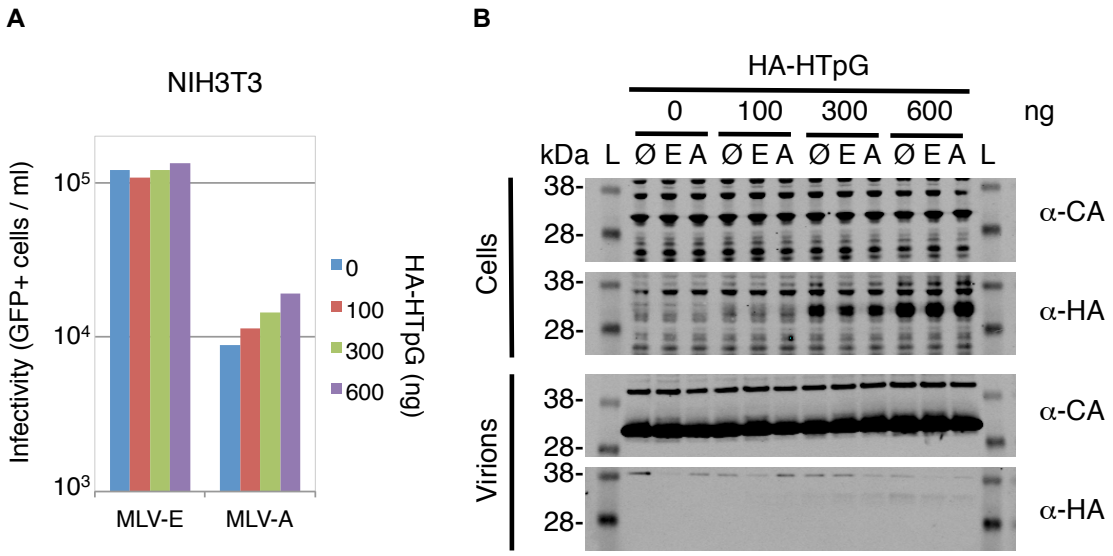


Figure 5.4: Effect of HTpG in MLV infectivity.

(A) Infectivity of MLV particles pseudotyped with ecotropic (MLV-E) or amphotropic (MLV-A) envelopes. Viruses expressing a GFP reporter protein were produced from 293T cells co-transfected with increasing amounts of a plasmid expressing HA-tagged HTpG or an empty vector. Viral titers were measured on NIH3T3 cells. **(B)** Western blot analyses of cell lysates and virions produced in (A). The cell lysates and virions were probed with anti-CA and anti-HA antibodies. L: molecular ladder. Ø: no envelope.

We then tested if the HTpG might have a positive effect on MLV particles pseudotyped with its own envelope (ancHTenv). For this purpose, and in order to directly test the activity of human SERINC5 on the ancestral envelope, we produced MLV particles pseudotyped with ancHTenv or the VSV glycoprotein (VSV-g), using 293T cells co-transfected with a plasmid expressing human SERINC5 and increasing amounts of the plasmid expressing HA-tagged HTpG, MLV glycoGag, SIV and HIV Nef or an empty vector. The produced viruses were then used to infect naïve 293T cells (Figure 5.6B). As expected human SERINC5 was not able to restrict MLV particles pseudotyped with VSV-g. However, MLV particles pseudotyped with ancHTenv were sensitive to human SERINC5 showing a reduced infectivity of almost 10 fold (Figure 5.6B). In agreement with previous reports the activity of human SERINC5 was blocked by the expression of SIV or HIV Nef and MLV glycoGag (Rosa et al., 2015; Usami et al., 2015) in a dose dependent manner (Figure 5.6B). Nevertheless, the expression of HTpG was not able to alleviate the inhibitory effect of human SERINC5 on ancHTenv pseudotyped MLV particles (Figure 5.6B), arguing against the functionality of HTpG as an analogous protein to MLV glycoGag.

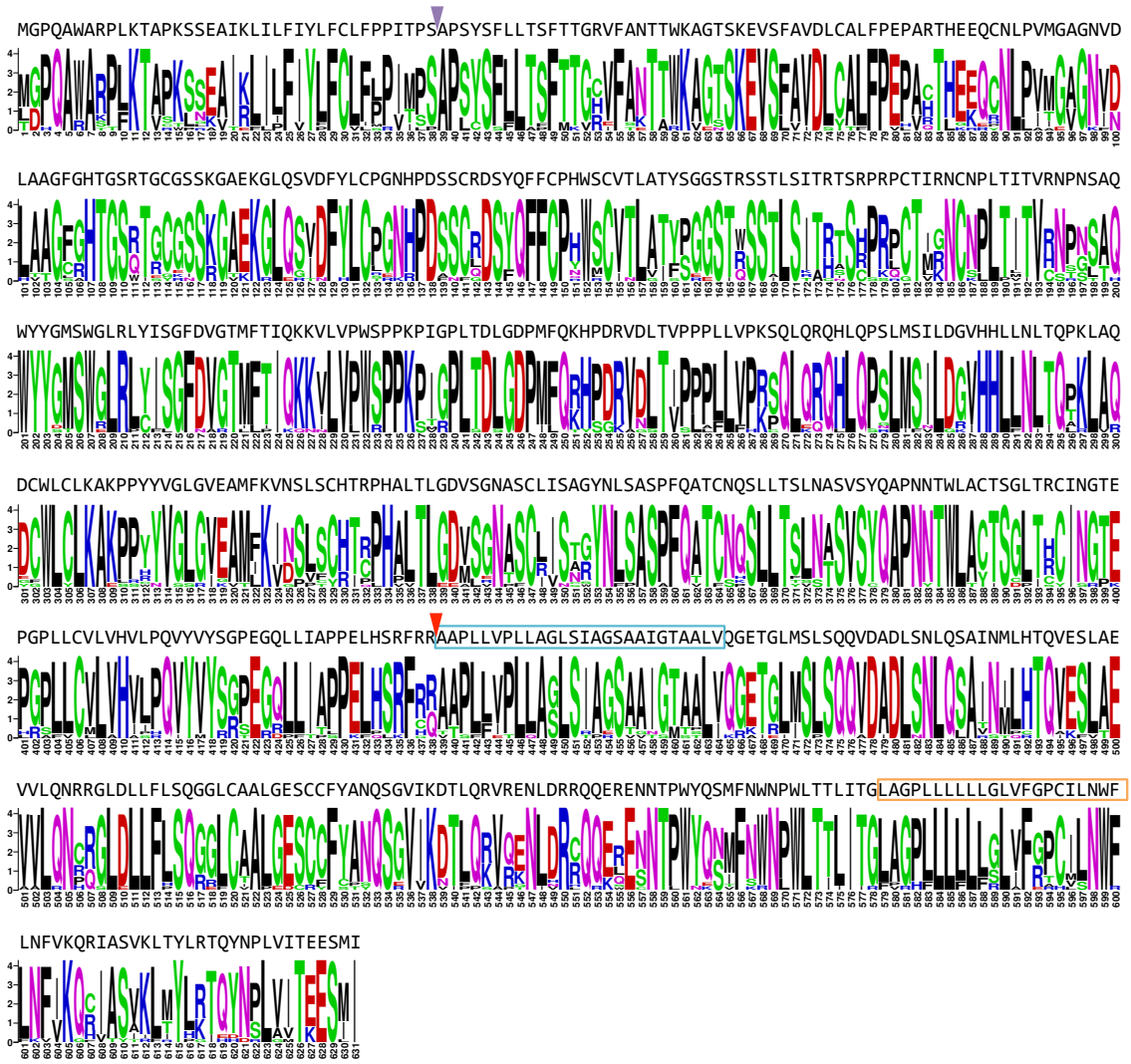


Figure 5.5: Refined ancestral HERV-T envelope sequence.

Logo plot of the translated *env* sequence in HERV-T3 proviruses. The total column height represents the sequence conservation on that site (in bits). The height of each residue represents its frequency at that particular site. The ancHTenv final sequence is showed on top of the logo plots. Signal peptide and propeptide furin cleavage sites are indicated by purple and red triangles, respectively. Sequence of the fusion peptide and transmembrane domain are indicted in cyan and orange boxes, respectively. Cleavage sites were predicted using ProP 1.0 (Duckert et. al. 2004).

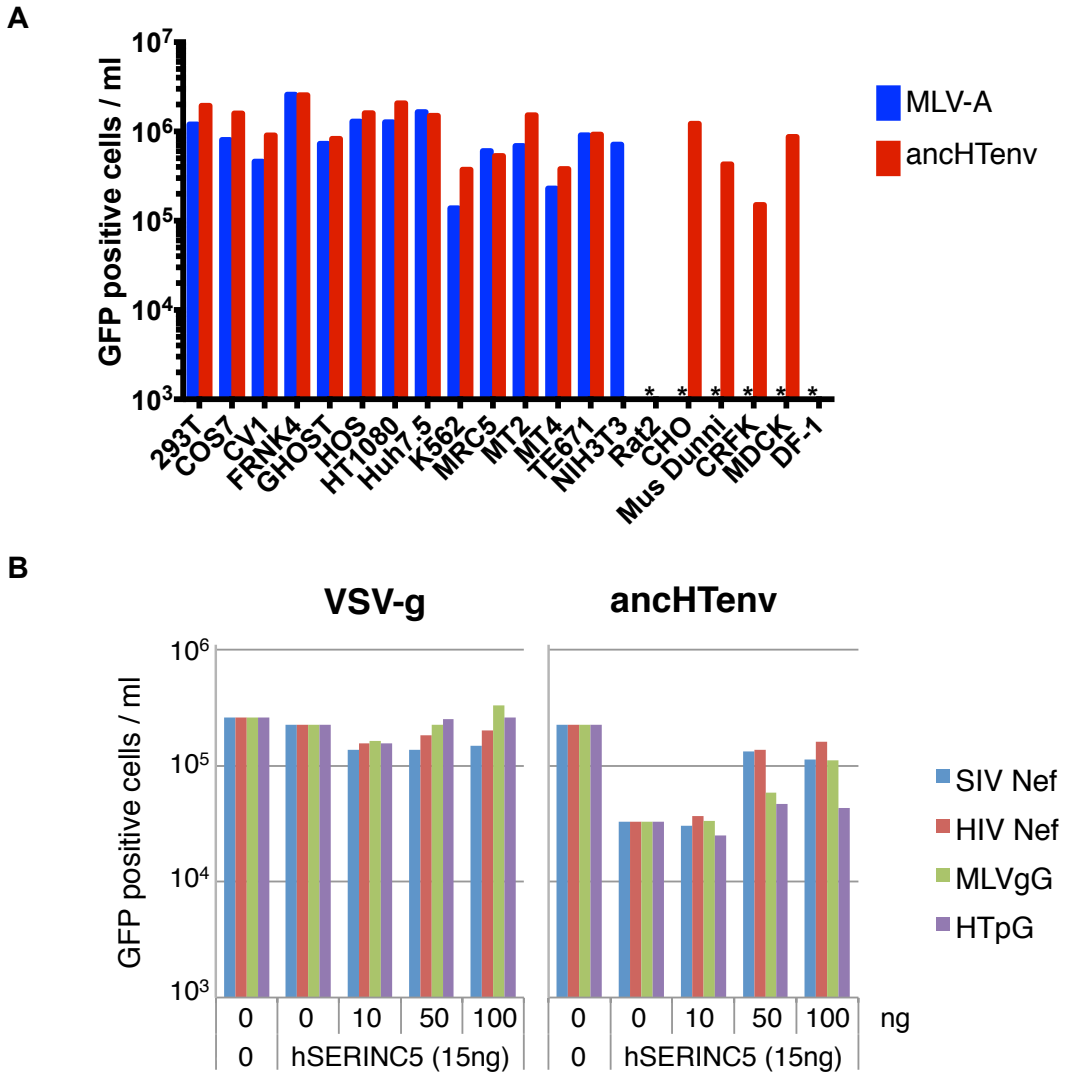


Figure 5.6: Infectivity of ancHTenv pseudotyped MLV particles and their inhibition by human SERINC5.

(A) Infectivity of MLV particles pseudotyped by ancHTenv or MLV-A in a variety of cell lines. Viruses expressing a GFP reporter protein were produced from 293T cells. (*): Not tested. **(B)** Effect of human SERINC5 on MLV particles pseudotyped by ancHTenv or VSV-g. Virus expressing a GFP reporter protein were produced from 293T cells co-transfected with 15ng of a plasmid expressing human SERINC5 and increasing amounts of plasmids expressing HTpG, MLV glycoGag, SIV and HIV Nef or an empty vector. Viral titers were measured on naïve 293T cells.

Existence of a ~17 MY old HERV-T3 env ORF in the human genome

As described in chapter I, although the vast majority of ERV sequences in animal genomes are highly mutated and inactivated, there are some examples of preserved (complete or partial) genes and structures that have been co-opted by the host as part of their genetic repertoire. A previous analysis for potentially complete ERV genes in the human genome revealed 16 protein coding *env* genes including one for HERV-T in chromosome 19 (de Parseval et al., 2003). This HERV-T envelope protein was mostly expressed in thyroid tissue with minor expression in prostate, kidney and other healthy tissues (de Parseval et al., 2003). Further studies by the same group showed that this envelope has lost its fusogenic property (Blaise et al., 2003). A closer look into this element in the human genome showed that this *env* gene was flanked by two HERV-T3 paired LTRs with inactivated *gag* and *pol* remains. A cDNA sequence is publicly available in GenBank (ID: AB266802) and indicates that the *env* mRNA is not generated from its 5' LTR, but from a HERV-L-like soloLTR (LTR66) located 3.5Kb upstream of HERV-T provirus. A search for orthologous loci to this particular provirus found its conservation in the genomes of gorillas and orangutans, but not in gibbons (Figure 5.7A), suggesting an integration date of ~13 to 20 MYA (Perelman et al., 2011). Whereas the human and orangutan sequences showed an *env* with full coding potential, the gorilla ORF is truncated due to a single nucleotide insertion at codon 234. This locus is also absent in chimpanzees due to a ~196Kb segmental deletion. The human HERV-T

envelope protein shares 85% identity (91% similarity) to the ancHTenv and its most obvious difference is the lack of the last five amino acids present at the C-terminus of ancHTenv. We cloned this human HERV-T3 *env* sequence (HsaHTenv) in an expression vector and tested its ability to pseudotype MLV particles. We also generated HA-tagged versions of ancHTenv, HsaHTenv and a codon-optimized version of HsaHTenv for expression in human cells (Figure 5.7B). MLV particles pseudotyped with ancHTenv were infectious, in accordance with previous experiments, with a slight reduction in infectivity for particles pseudotyped with a HA-tagged version of this envelope. On the contrary none of the constructs expressing HsaHTenv were able to pseudotype MLV particles and produce infection (Figure 5.7B). Western blot analyses of producer cells and virus lysates clearly showed a defect in the processing of HsaHTenv, compared to the ancHTenv, that prevents incorporation of a ~25kDa protein fragment into virions (Figure 5.7C), most likely corresponding to the TM protein domain predicted to weight 23.92kDa (Figures 5.5) (see chapter II). As expected the codon-optimized version of HsaHTenv is expressed at higher levels than the sequence cloned form human gDNA (Figure 5.7C). For this reason all the experiments referring to HsaHTenv from this point forward will be performed using the codon-optimized version of this envelope.

Figure 5.7: Conservation of a human HERV-T3 *env* gene with coding potential and its infectivity.

(A) Graphical representation of a glocal sequence alignment of a human HERV-T *env* and its orthologs in gorillas and orangutans generated using mVISTA (Frazer et al., 2004). Each graph shows the percentage of conservation between the DNA sequences of that organism with the corresponding human sequence at any given coordinate. Sequences that show more than 90% similarity with the human sequence are indicated in color. Non-coding regions are indicated in pink. Coding *env* sequences are indicated in purple, HERV-T3 proviral sequence is indicated in light blue. The location of the LTR66 promoter is also indicated. **(B)** Infectivity of MLV particles pseudotyped with HA tagged and untagged MLV-A, ancHTenv or the human HERV-T envelope (HsaHTenv). (opti-) indicates a codon optimized version for expression in human cells. Viruses expressing a GFP reporter protein were produced from 293T cells. Viral titers were measured on 293T and HT1080 cells. Data from 3 independent experiments. **(C)** Western blot analyses of cell lysates and virions produced in (B). The cell lysates and virions were probed with anti-CA and anti-HA antibodies. L: molecular ladder. Ø: no envelope.

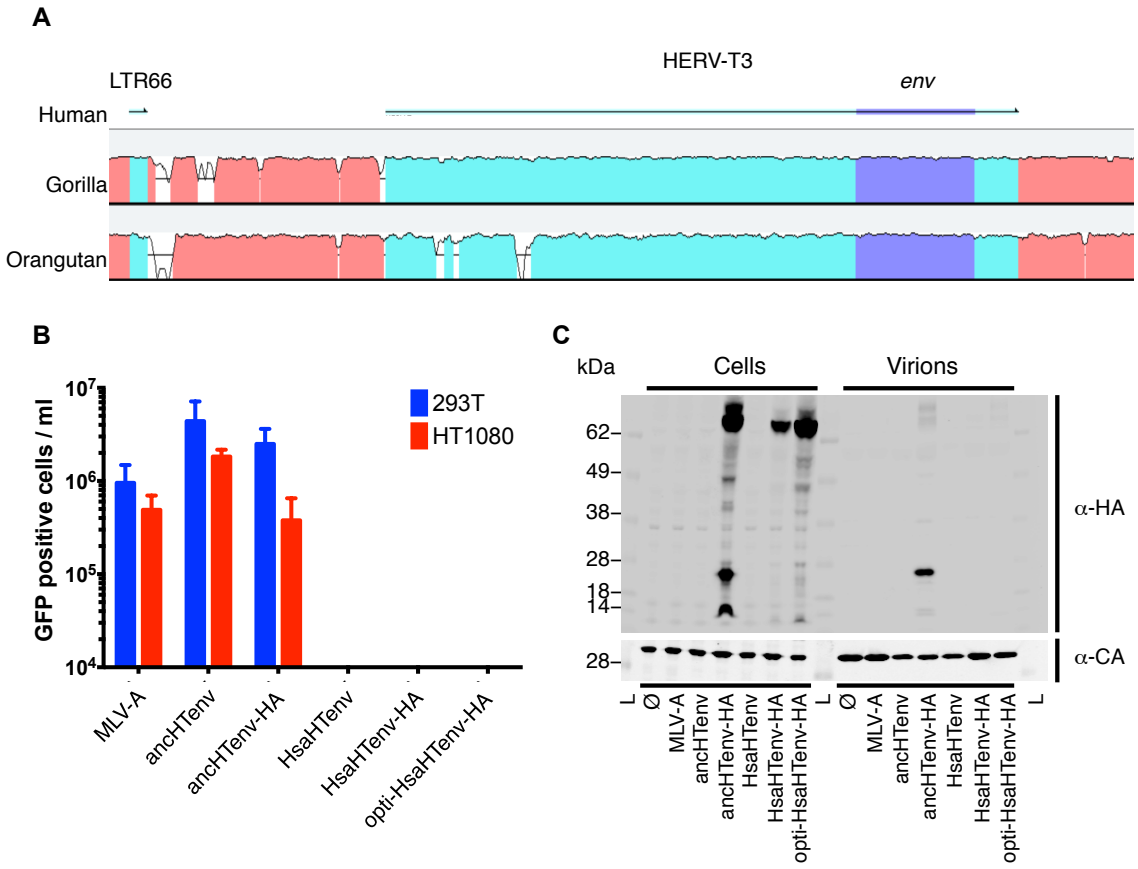


Figure 5.7: Conservation of a human HERV-T3 env gene with coding potential and its infectivity.

The failure of HsaHTenv to produce infectious particles despite its coding potential, made us speculate that perhaps an ancestral sequence for this particular locus would revert the inactivating mutations and produce infectious particles. The reduced set of sequences (only three sequences from human, gorilla and orangutan) prevents us from using previously described ML methodologies. Instead, we produced an ancestral sequence for this locus by selecting residues based on the phylogenetic relationships of the three sequences. Particularly polymorphic sites were resolved by comparison to the ancHTenv sequence. The resulting ancestral sequence for this particular locus (ancHsaHTenv) was codon-optimized, synthesized and cloned into an expression vector to test its infectivity. Similarly to HsaHTenv, MLV pseudotyped particles with ancHsaHTenv were non-infectious and showed similar defects in processing and virion incorporation compared to ancHTenv (Figures 5.8B and C). Comparisons of the envelope protein sequences revealed the presence of a mutated and non-functional furin cleavage site in HsaHTenv and their orthologous sequences in gorilla and orangutan (Figures 5.5 and 5.8A), possibly explaining the defects in processing and incorporation into virions. In order to construct a furin site mutant ancHTenv and an HsaHTenv and ancHsaHTenv with the furin site repaired, we switched the furin cleave motifs between ancHTenv and both HsaHTenv and ancHsaHTenv. These constructs were used to pseudotype MLV particles and test their processing and infectivity. The furin site mutation introduced in ancHTenv was sufficient to impair the production of

infectious particles, the correct processing of the envelope protein and the incorporation of the TM domain into virions (Figures 5.8B and C). Strikingly, neither HsaHTenv nor ancHsaHTenv with the furin sites repaired were able to produce infectious particles (Figure 5.8B). The repair of the furin cleavage site of HsaHTenv also failed to produce the correct signal corresponding to the TM domain and instead resulted in the emergence of higher molecular weight fragments (Figure 5.8C). However, the same modification in ancHsaHTenv was able to produce a similar processing pattern as the functional ancHTenv and the presence of a fragment corresponding to the molecular weight of the TM domain (Figure 5.8C). This fragment was also incorporated into viral particles, although to a lesser extent than the one produced by ancHTenv (Figure 5.8C). Additionally, only the expression of ancHTenv, but no other HERV-T envelopes, results in the production of syncytia in 293T cells (Figure 5.9), supporting the fusogenic potential of this ancestral envelope and confirming its absence in the rest of the HERV-T envelopes tested. These results indicate that the complete protein-coding HERV-T env present in the human genome is not able to serve as a functional envelope protein and an alternative function should be further assessed.

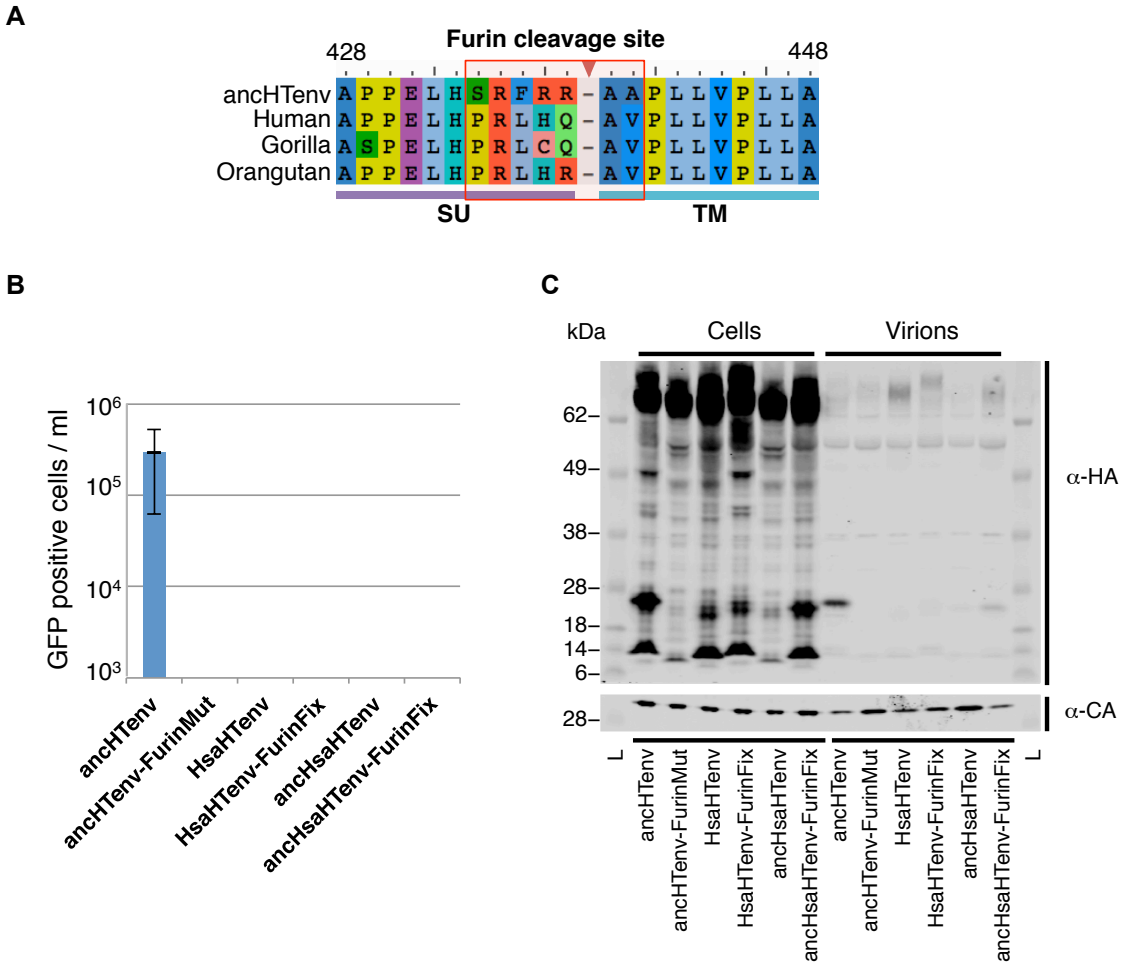


Figure 5.8: Infectivity and processing of HERV-T envelopes.

(A) Multiple sequence alignment of ancHTenv and HERV-T envelopes in primate genomes. The propeptide furin cleavage site is indicated by a red box and triangle. Numbers above the alignment indicate amino acid positions. (B) Infectivity of MLV particles pseudotyped with HA tagged ancHTenv, HsaHTenv, and a recent ancestral sequence of the human protein-coding *env* loci (ancHsaHTenv) with or without a predicted functional furin site (FurinFix, FurinMut). Viruses expressing a GFP reporter protein were produced from 293T cells. Viral titers were also measured on 293T cells. Data from 3 independent experiments. (C) Western blot analyses of cell lysates and virions produced in (B). The cell lysates and virions were probed with anti-CA and anti-HA antibodies. L: molecular ladder.

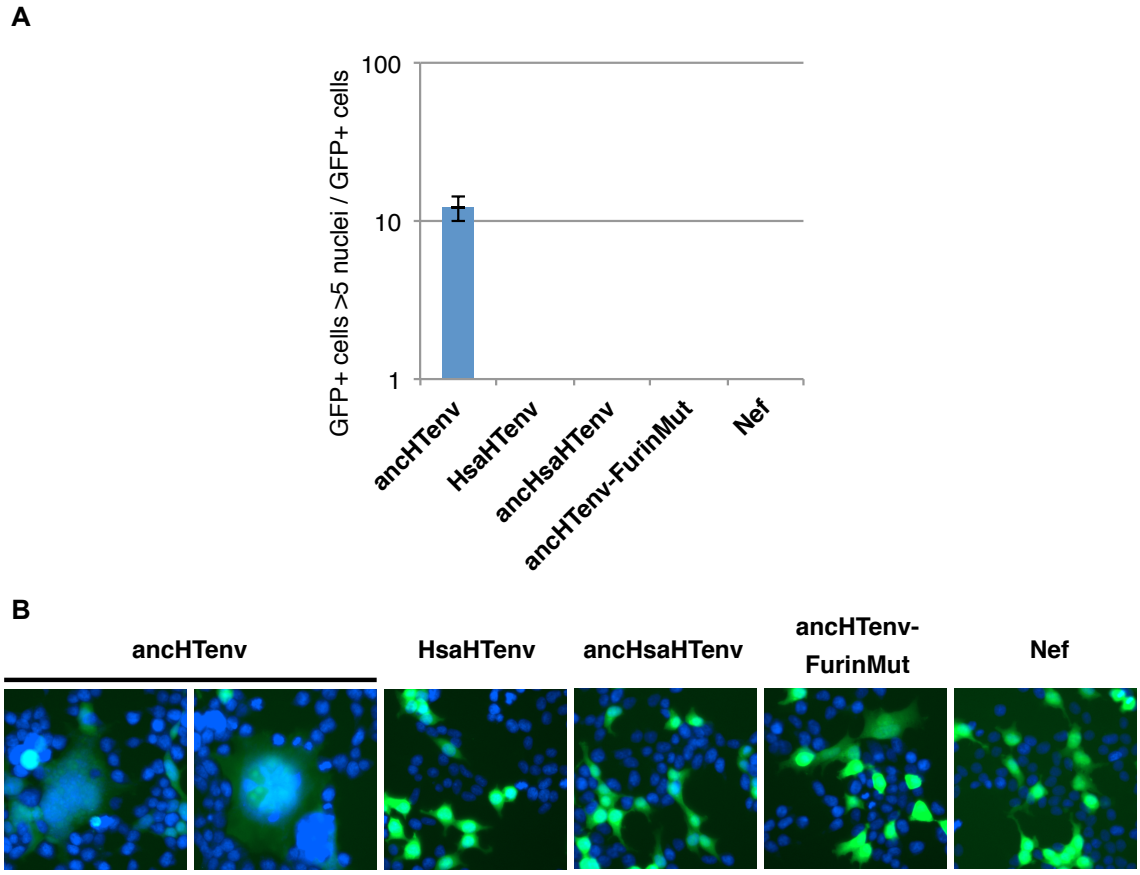


Figure 5.9: Fusogenic properties of different HERV-T envelopes.

(A) Ratio of cells with more than 5 nuclei over the total number of cells expressing the different HERV-T envelopes. Data calculated from 30 frames of 293T cell lines stably expressing different HERV-T envelopes. Envelope expression correlates with the expression of a GFP reporter gene included in the lentiviral vector used to produce the stable cell lines. **(B)** Example of multinucleated cells quantified in (A). Fixed 293T cells were incubated with anti-HA antibodies. Nuclear DNA was labeled with DAPI. Images were amplified and centered on features of interest.

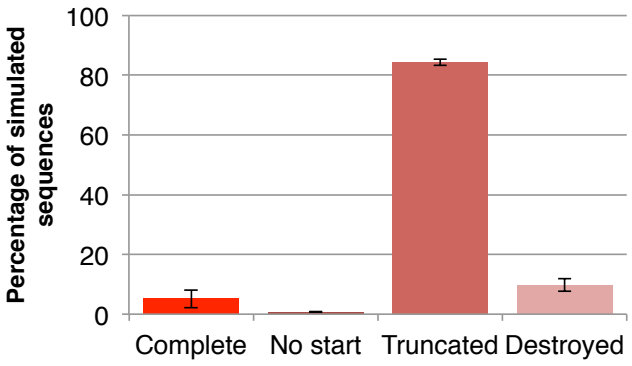
Selective pressures have conserved the coding potential of HsaHTenv.

The inability of HsaHTenv to produce infectious particles despite coding potential, made us re-evaluate the significance of finding such a complete *env* gene in the human genome. We first aimed to address how likely is to find a complete *env* ORF assuming there are no selective pressures to conserve any function (i.e. Neutral evolution). For this purpose we performed Monte-Carlo simulations of *in silico* neutral evolution on the ancHsaHTenv sequence for 13.45 and 19.68 MY (the minimum and maximum estimates of the origin of hominids (Perelman et al., 2011)). On average, only 5% of all simulated sequences (2–8% of 10,000 iterations) retain coding potential showing no truncation of its ORF and a 5' methionine codon (Figure 5.10A). The majority of the sequences (83.3–85.5%) retained the 5' methionine but they were truncated by the presence of at least 1 stop codon (Figure 5.10A). The distribution of stop codons among the simulated sequences had a median of 2–3 with 70–86% of the sequences showing at least 2 stop codons. The results of these simulations are probably an underestimate because we are not taking in to account the additional inactivating mutations derived from insertion/deletion processes introduced by genetic drift. Therefore it is unlikely that the preserved *env* ORF present in the human genome is the result of ~17 MY of evolution under no selective pressures.

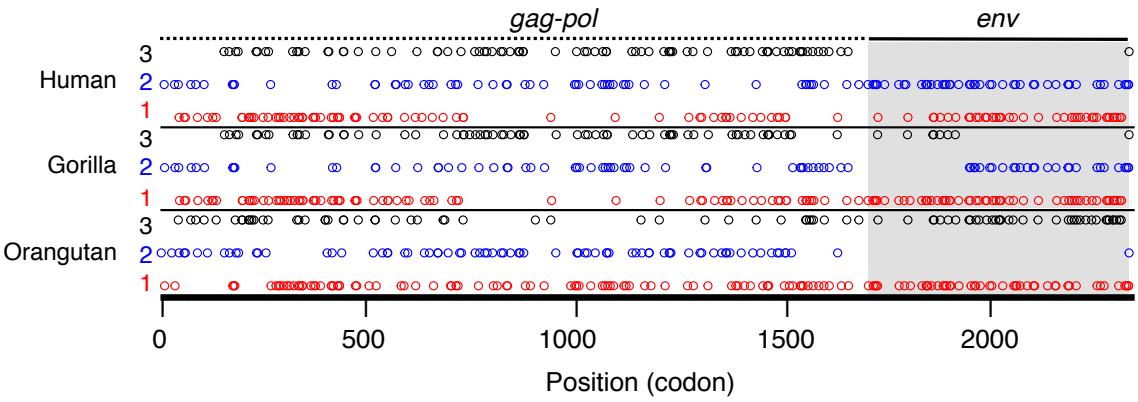
Figure 5.10: Analysis of protein-coding HERV-T *env* sequences under neutral evolution.

(A) Results of the Monte-Carlo simulations of *in silico* neutral evolution. The ancHsaHTenv nucleotide sequence was evolved for 13.45 (minimum estimate) and 19.68 (maximum estimate) MY under a human neutral substitution rate of $2.2x10^{-9}$ substitutions per site per year (Lander et al., 2001). 10,000 simulated sequences were generated using seq-gen (Rambaut and Grassly, 1997). Sequences were classified as: Complete: showing 5' Methionine and no stop codons. No start: absence of 5' Methionine. Truncated: presence of 1 or more stop codons. Destroyed: absence of 5' Methionine and presence of 1 or more stop codons. The average values are plotted with maximum and minimum estimates indicated by error bars. **(B)** Distribution of stop codons along the coding-sequence of HERV-T proviruses present in human, gorilla and orangutan. Stop codons present in the forward 1^{st}, 2^{nd} and 3^{rd} frames are indicated by red, blue and black circles, respectively. The location of the *env* coding sequence is colored in gray. **(C)** Estimation of speciation dates was performed separately for *gag-pol* and *env* sequences by adjusting the divergence of any given sequence to each other under a human neutral substitution rate. (*) As comparison, maximum and minimum estimations of the speciation dates were obtained from Perelman et al., 2011. A cladogram with the phylogenetic relations between the species is shown on the left. The numbers in the cladogram indicate internal nodes corresponding to speciation events.

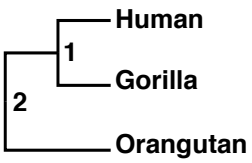
A



B



C



Node	Speciation date (MY) max*	min*	*gag-pol*	*env*
1	10.07	6.58	5.30	4.95
2	19.68	13.45	13.55	9.93

Figure 5.10: Analysis of protein-coding HERV-T *env* sequences under neutral evolution.

While the *env* gene has remained complete in humans and orangutans, the *gag* and *pol* genes show inactivating mutations suggestive of neutral evolution. In fact, by mapping the existence of stop codons along the entire proviral coding sequence, it is strikingly obvious that the region where *env* is present has been refractory to non-sense mutations, compared with the gag-pol region, in all three species (Figure 5.10B). It is worth noting that despite the truncation of the gorilla *env* ORF, the remaining *env* sequence is preserved in a different frame (frames 2 and 3 of the gorilla sequence) (Figure 5.10B). We decided to exploit this discrepancy and approximate the speciation dates of hominids utilizing both types of sequences separately. The estimated speciation dates were calculated by adjusting the divergence of any given sequence to each other under a human neutral substitution rate of 2.2×10^{-9} substitutions per site per year (Lander et al., 2001) (Figure 5.10C). Whereas the putative speciation dates obtained from *gag-pol* sequences were almost equal to the minimum previous estimates (5.30 and 13.55 MYA for human-gorilla and orangutan speciation events, respectively) (Perelman et al., 2011) (Figure 5.10C), the approximated dates obtained using the *env* sequences were smaller (4.95 and 9.93 MYA human-gorilla and orangutan speciation events, respectively) (Figure 5.10C). Although the discrepancy between the dates obtained using distinct sequences of the same locus is small, this difference suggests that the *env* sequence is evolving slower than the rest of the provirus. This observation, together with the results of the simulation and the resistance to non-sense mutations, point out for the existence

of selective pressures that have acted on this locus to preserve its coding capacity. The completeness of the *env* ORF perhaps reflects its co-option by an ancestral hominid to perform an unknown, but critical function and therefore was fixed in the ancestral population ~17 MYA.

The HsaHTenv is a potent inhibitor of particles pseudotyped with ancHTenv.

One intriguing possibility is that the protein-coding *env* copy was co-opted to block infection by HERV-T or related incoming viruses. In order to test this hypothesis we transduced 293T cells with lentiviral vectors containing the different HERV-T envelope constructs together with a GFP reporter gene to monitor their expression. These cells were then challenged with MLV particles pseudotyped with ancHTenv that express a RFP reporter gene upon infection. The resulting fluorescent cell populations were quantified by fluorescence-activated cell sorting (FACS). As expected, ancHTenv expression is able to potently block infection of MLV particles pseudotyped with the same envelope (ancHTenv) showing a ~300 fold reduction in infectivity compared to cells transduced with non-relevant proteins (HIV-1 Nef and GFP) (Figure 5.11). The effect on infectivity is slightly reduced upon mutation of the Env furin cleavage site, showing that the efficient processing of the env polyprotein is not necessary for its inhibitory effect (Figure 5.11). In accordance with this finding the non-functional complete envelope present in humans and its most recent ancestor (HsaHTenv and ancHsaHTenv) showed a 13 and 35 fold inhibition of MLV

particles pseudotyped with the functional ancHTenv, respectively, when compared to non-relevant proteins (Figure 5.11). These results argue for a possible antiviral function of the human complete *env* gene (HsaHTenv) that was preserved for ~17 MY.

To further corroborate these findings, we produced single cell clones of 293T cells stably expressing distinct HERV-T envelopes and characterized their ability to be infected by MLV particles pseudotyped with ancHTenv. Although single cell clones were selected using an antibiotic resistant marker, there were subpopulations of cells showing distinct sensitivity to ancHTenv virus (Figure 5.12). In order to correctly address the expression of the different HERV-T envelopes, we produced single cell clones stably expressing HA-tagged versions of the HERV-T envelopes and quantified their expression by immunofluorescence (Figure 5.12D). After classifying each single cell clone into four discrete categories depending on the expression of its correspondent envelope (Figure 5.12D), we challenged them with MLV particles pseudotyped with ancHTenv (Figure 5.12A, B and C). This classification resulted in a homogeneous sensitivity to ancHTenv viruses, that is inversely correlated with the percentage of cells expressing the HERV-T envelopes, showing on average a ~10 fold reduction in infectivity by the highest expressing cells. Remarkably, the antiviral effect of the HERV-T envelopes is specific to viruses pseudotyped with ancHTenv having no effect on viruses pseudotyped with a different envelope (MLV-A). In summary the previous experiments indicate that despite their failure

as functional envelopes, HsaHTenv and ancHsaHTenv show a specific inhibitory activity against viruses pseudotyped with ancHTenv, suggesting that this antiviral function might have been exploited by an ancestral hominid leading to the fixation of the proviral locus and the preservation of the coding capacity of the *env* gene.

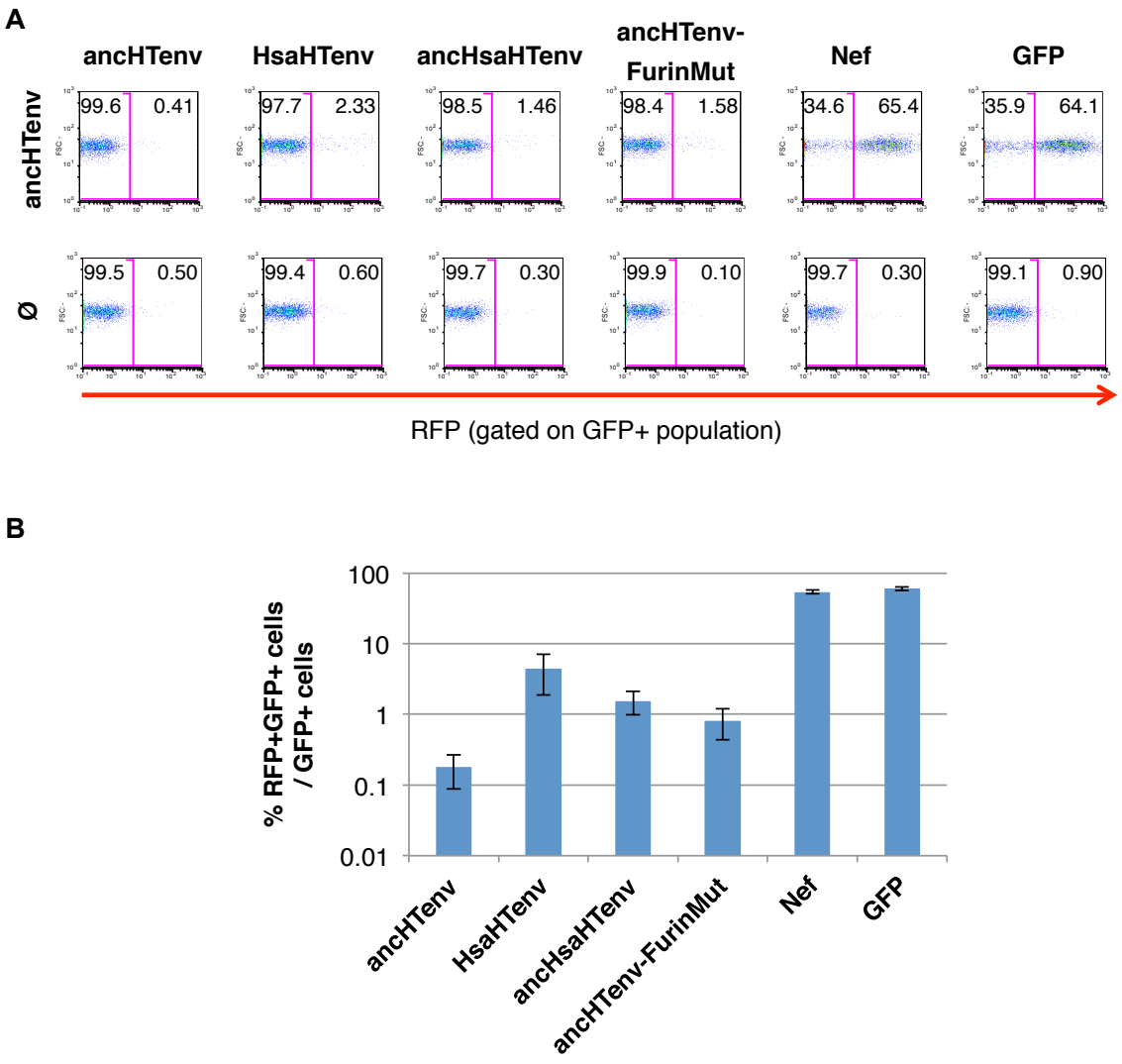


Figure 5.11: Antiviral capacity of different HERV-T envelopes.

(A) FACS plots showing the inhibitory effect of HERV-T envelopes. 293T cells were transduced with lentiviral vectors expressing ancHTenv, HsaHTenv, ancHsaHTenv, ancHTenv-FurinMut, HIV-1 Nef and GFP together with a GFP reporter gene. Cells were challenged with concentrated MLV particles pseudotyped with ancHTenv that express RFP upon infection. Plots describe the percentage of RFP positive cells gated on the GFP positive cell population. Ø: no virus. **(B)** Percentage of RFP and GFP positive cells over total GFP positive population. Data of 3 individual experiments performed as in (A).

Figure 5.12: The antiviral effect of different HERV-T envelopes is specific to particles pseudotyped with ancHTenv and is dependent on their expression.

293T single cell clones stably expressing HA-tagged or un-tagged versions of **(A)** ancHTenv, **(B)** HsaHTenv or **(C)** ancHsaHTenv were infected with concentrated MLV particles pseudotyped with ancHTenv or MLV-A and expressing a GFP reporter protein. Single cell clones expressing HA-tagged versions of the HERV-T envelope proteins were classified depending on the percentage of cells expressing the different envelopes. **(D)** Criteria for classification of HA-tagged proteins in (A, B and C). Single cell clones stably expressing the different HERV-T envelopes were fixed and incubated with anti-HA antibodies. Nuclear DNA was labeled with DAPI.

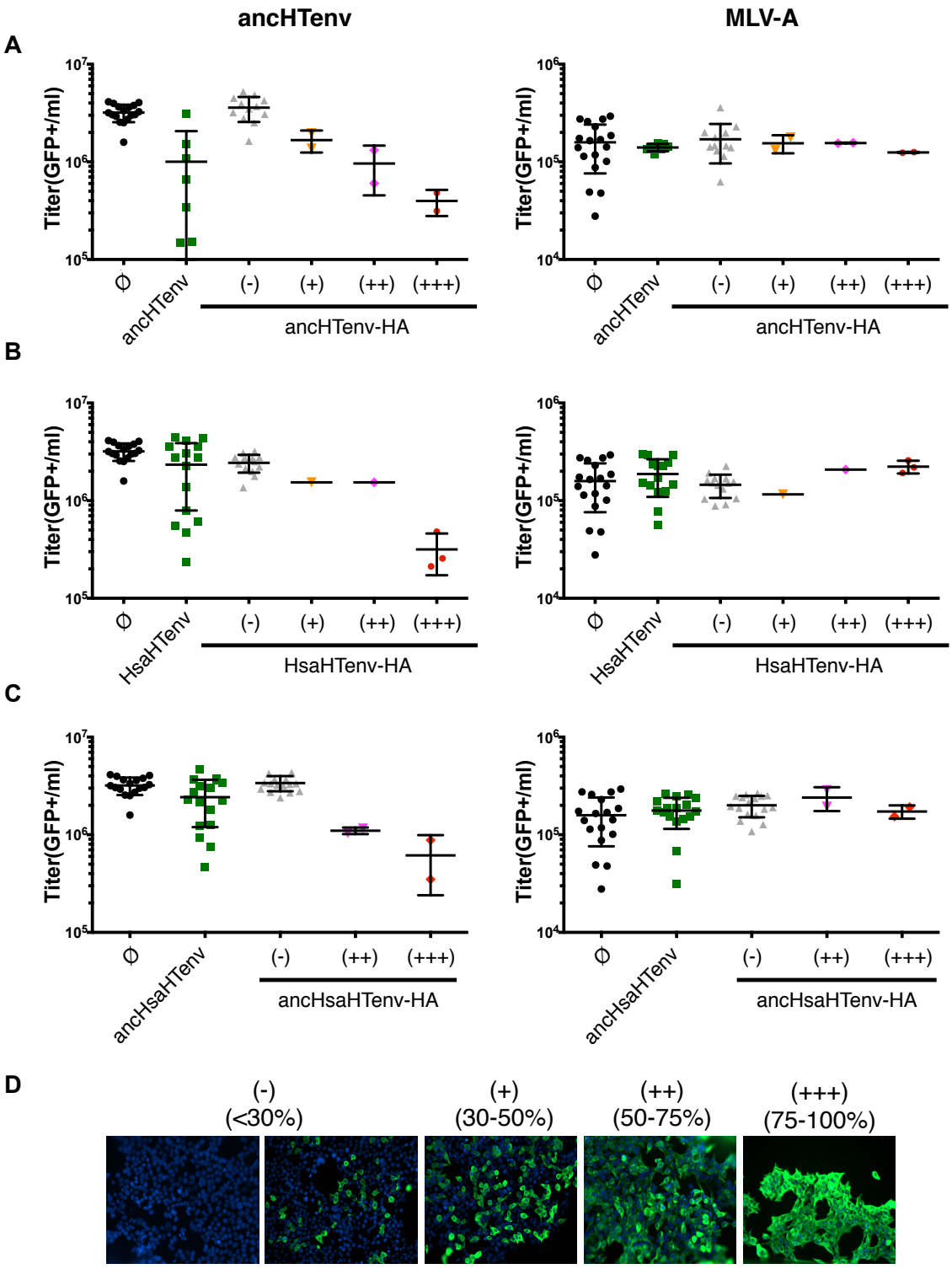


Figure 5.12: The antiviral effect of different HERV-T envelopes is specific to particles pseudotyped with ancHTenv and is dependent on their expression.

Identification of the receptor used by ancHTenv

A probable mechanism by which HsaHTenv and ancHsaHTenv restrict infection by ancHTenv viruses could be by blocking and sequestering the available receptor molecules used by ancHTenv to infect cells. In order to test this hypothesis we first needed to determine the identity of this receptor. For this purpose we created a lentiviral library expressing mRNAs cloned from cDNA molecules from a permissible cell line (293T), together with an antibiotic resistance selection marker (blasticidin). The resulting library contained a complexity of 3.5 x 10^6 colony forming units (cfu), suggesting a good representation of the different cDNA molecules. We first tested the infectivity of our cDNA library in cells that were not sensitive to ancHTenv pseudotyped particles (Figure 5.6A). Viruses containing the cDNA library were 10 fold more infectious in DF-1 cells than NIH3T3 (3.7 x 10^5 cfu/ml and 3 x 10^6 cfu/ml, respectively), therefore we chose to perform the screening for the ancHTenv receptor in DF-1 cells.

Preliminary infections of DF-1 cells with MLV particles pseudotyped with ancHTenv containing a *neo* resistant gene revealed a minor, but reproducible, infection in this cells with titers of ~4 x 10^2 G418-resistant colonies/ml. Intrigued by this finding we also infect NIH3T3 cells with the same virus pseudotyped with ancHTenv and found that these cells were also sensitive to ancHTenv viruses with titers of ~1 x 10^3 G418-resistant colonies/ml. It is possible that the level of infection showed on these cells were below the detection power of previous

experiments using GFP as a reporter gene, such that the few infected cells needed to expand in order to acquire the significant levels required for detection, levels that were achieved by 10 days in antibiotic selection.

This “background” infection in DF-1 cells complicated our efforts to perform the receptor screening. To overcome this difficulty we decided to challenge and select DF-1 cells expressing the cDNA library with ancHTenv pseudotyped MLV particles expressing different fluorescence and antibiotic resistant genes iteratively. The rationale was that every time we challenge cDNA expressing DF-1 cells and select infected cells for another challenge, we would be sequentially increasing the population of “true receptor” expressing DF-1 cells over the population of “irrelevant cDNA” expressing cells (selected at random due to background infection (Figure 5.13A)). Following this strategy we infected DF-1 cells with cDNA library containing lentiviruses at an MOI of 8, to increase the chance that cells will express the “true receptor” regardless of its representation in the library. Two days after the initial infection, cDNA expressing DF-1 cells were challenged with MLV particles pseudotyped with ancHTenv containing a *neo* resistant gene. Cells were placed in G418 selection two days post infection (dpi) and resistant colonies were collected after 10 days. G418 resistant DF-1 cells were then challenged with ancHTenv pseudotyped MLV particles containing a hygromycin resistant gene. Cells were placed in hygromycin selection two dpi and resistant colonies were collected after 10 days (Figure 5.12A). At this point, although the number of hygromycin resistant colonies was low, they were

sufficiently higher than background infection observed for naïve DF-1 control cells, using the same virus. Finally, hygromycin resistant DF-1 cells were infected with ancHTenv pseudotyped MLV particles expressing RFP upon infection (Figure 5.13A). Almost 18% of the hygromycin resistant DF-1 cells were also RFP positive, compared to less than 1% for the naïve DF-1 control cells. Next, we extracted gDNA from three RFP positive cell populations and amplified possible receptor candidates by PCR (with primers designed to anneal to the lentiviral vector) (Figure 5.13B). Surprisingly only two products were amplified from gDNA, the strongest band corresponded to a non-coding RNA (ncRNA) derived from the human autophagy related 12 (*atg12*) gene (Figure 5.13B). This strong signal might be due to an overrepresentation of this ncRNA in the cDNA library reflecting their endogenous levels in 293T cells. Alternatively its size and composition might have out-competed other PCR substrates present. The second, and much weaker ~1.8Kb band corresponded to the human monocarboxylate transporter 1 (MOT1, also known as SLC16A1) a multi-transmembrane protein that mediates the transport of monocarboxylates such as lactate and pyruvate across the plasma membrane (Halestrap and Wilson, 2012) (Figure 5.13B). Given the propensity for gammaretroviruses to utilize multi-transmembrane solute transporters as their receptors (Tailor et al., 2003) (Figure 1.4A), MOT1 was a strong candidate to be the receptor used by ancHTenv.

To validate the function of MOT1 as the receptor once used by HERV-T, we cloned the amplified MOT1 sequence (Figure 5.13B) back into a lentiviral vector

to make DF-1 cells stably expressing human MOT1. MLV particles pseudotyped with ancHTenv, expressing GFP or RFP upon infection, are able to infect exclusively DF-1 cells that stably express human MOT1 but not naïve DF-1 cells or those stably expressing GFP (Figure 5.14). These results confirm that MOT1 is indeed the receptor used by ancHTenv and probably used initially by HERV-T ~34 MYA.

Figure 5.13: ancHTenv receptor screening using a cDNA library derived from 293T mRNA.

(A) Scheme of the receptor screening strategy. DF-1 cells were infected with the cDNA lentiviral library. Two dpi cells were challenge with an ancHTenv pseudotyped MLV particles containing a *neo* gene. Two dpi cells were placed in G418 selection for 10 days. G418 resistant cells were challenged with ancHTenv pseudotyped MLV particles containing a hygromycin resistance gene. Two dpi cells were placed in hygromycin selection for 10 days. Hygromycin resistant cells were challenged with an ancHTenv pseudotyped MLV particles expressing RFP upon infection. Two dpi cells were analyzed for the expression of RFP. gDNA was extracted from RFP positive cells and receptor candidates were amplified by PCR using primers that annealed to the lentiviral vector. Virus and cells expressing the cDNA of the ancHTenv receptor or other cDNA are colored in cyan and red, respectively. The contour of the cell colonies represents the expression of a particular reporter gene. Cell plates on the right represent the control DF-1 cells that were not infected with the lentiviral library. **(B)** Principal products of the PCR amplification in the last step of (A). The bands corresponding to the *atg12* non-coding RNA and MOT1 are indicated. Image from gDNA PCR from 3 RFP positive DF-1 cell populations.

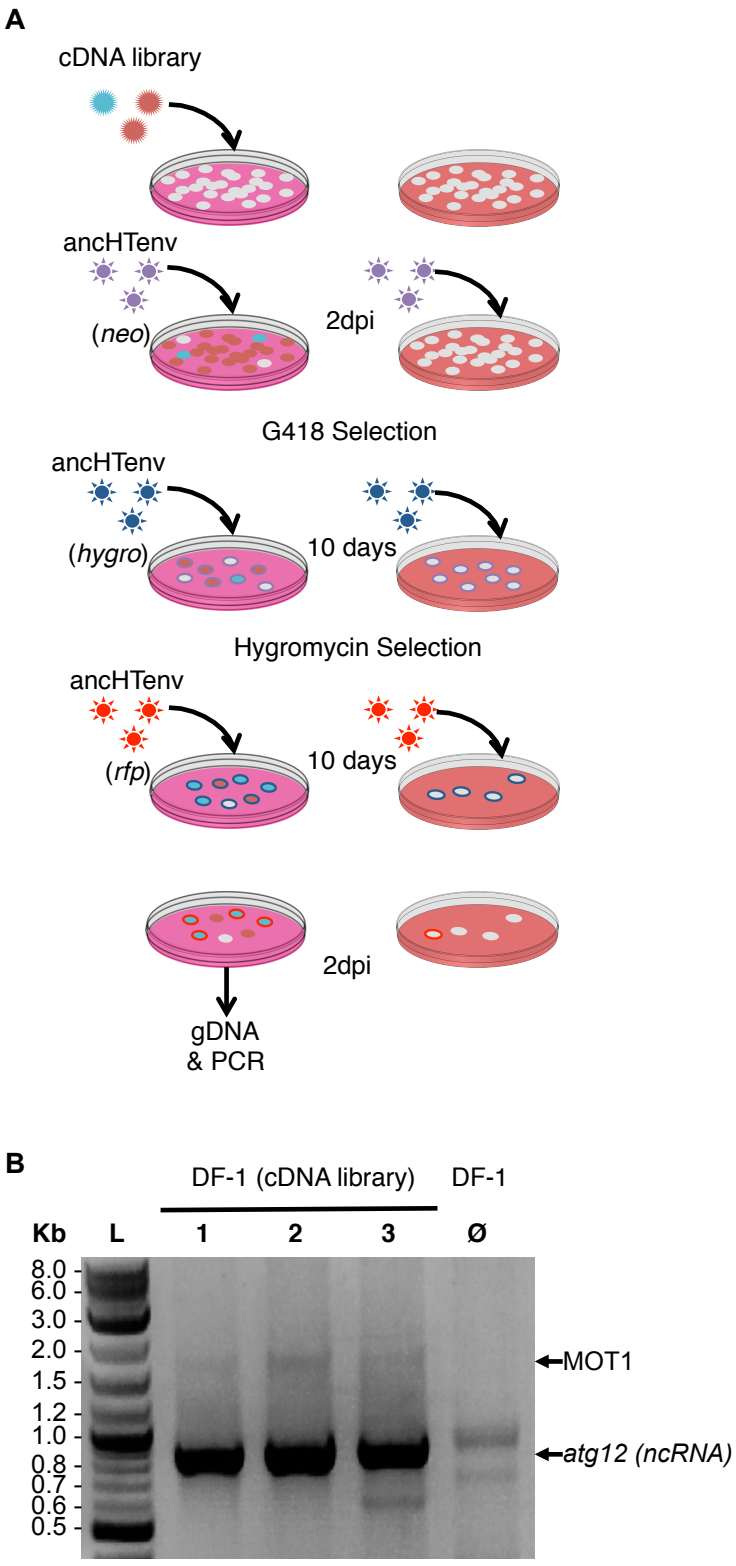


Figure 5.13: ancHTenv receptor screening using a cDNA library derived from 293T mRNA.

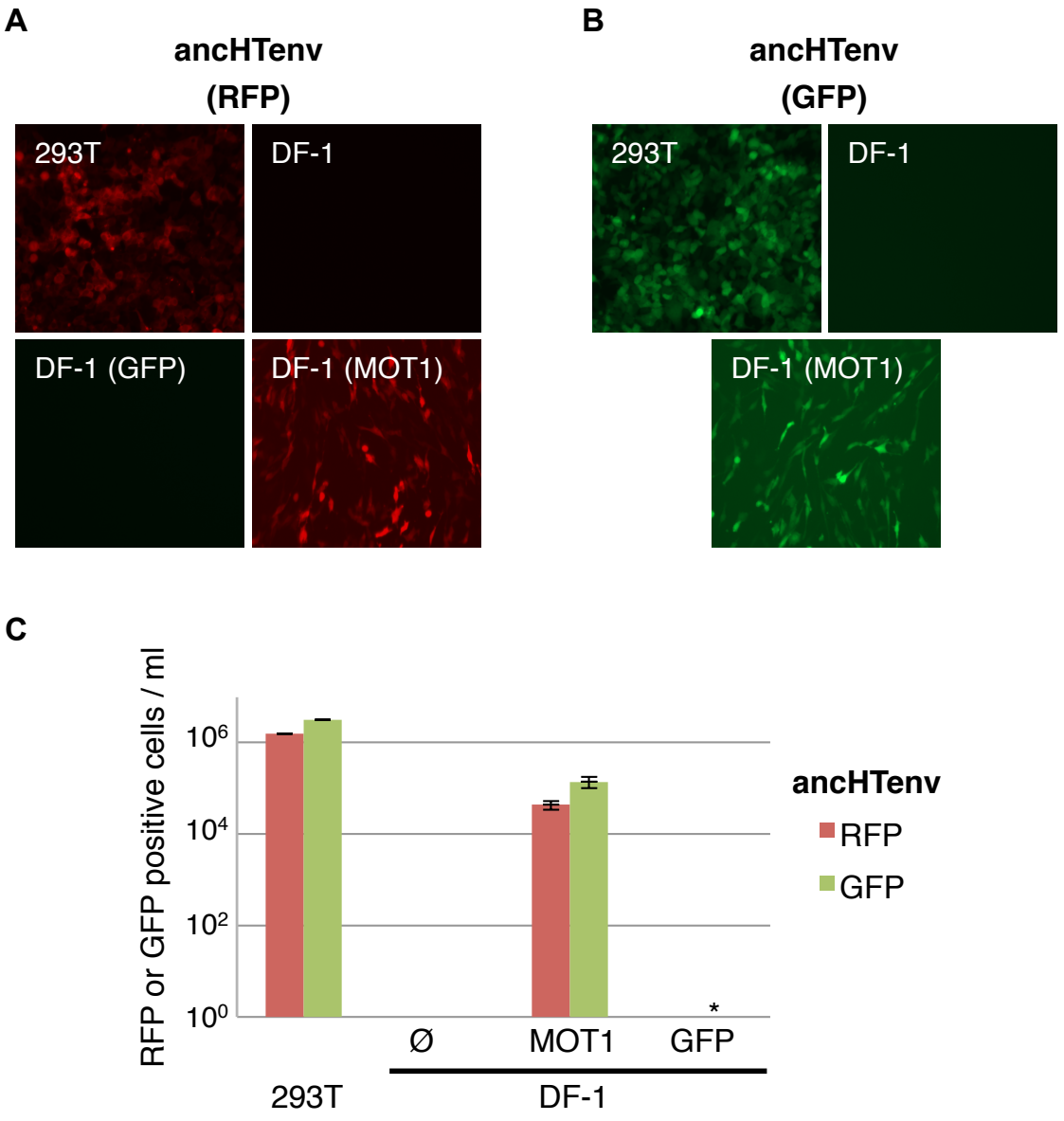


Figure 5.14: Validation of MOT1 as the ancHTenv receptor.

(A-B) Fluorescent microscopy of 293T, naïve DF-1 or DF-1 cells stably expressing GFP or MOT1 infected with MLV particles pseudotyped with ancHTenv and expressing (A) RFP or (B) GFP upon infection. **(C)** Viral titers of the infections in (A-B). Data from 2 independent cell populations. (*): Not tested.

Summary

In this chapter we reconstructed the ancestral genomic sequence of HERV-T, a low copy number primate ERV that is closely related to gammaretroviruses. By analyzing the unusually long ancestral HERV-T3 leader sequence we were able to find an 855nt ORF separated from gag by 36nt. This *pre-gag* ORF of unknown function putatively codes for a protein that includes a transmembrane domain. Additional analysis of the HERV-T3 ancestral sequence allowed us to estimate the corresponding ancestral envelope protein sequence (ancHTenv). We found that a modern gammaretrovirus (MLV) could be pseudotyped with ancHTenv enabling it to infect a wide variety of primate cell lines with titers that are similar to MLV particles carrying the amphotropic MLV envelope. A single HERV-T proviral insertion in the genome of all great apes contains an *env* gene with full coding potential. Proteins encoded by the extant human HERV-T *env* gene (HsaHTenv) and one estimated to be encoded by the hominid ancestor were not able to generate infectious MLV pseudotyped particles, probably because HsaHTenv is not correctly processed into its mature, functional form. Statistical and phylogenetic analyses indicate that the *env* gene in this locus is evolving slower than the rest of the proviral sequences, and its coding capacity cannot be explained by genetic drift, suggesting that selective pressures have acted on this locus to conserve its envelope sequence. Remarkably, we found that expression of the extant HsaHTenv was able to block infection by MLV particles pseudotyped with the ancHTenv, but not particles pseudotyped with the

amphotropic MLV envelope. These results suggest a potential role of HsaHTenv as a restriction factor. Additionally, we identified MOT1 as the receptor used by ancHTenv. However, further experiments are needed in order to test the hypothesis that HsaHTenv served as a restriction factor through the interference with the receptor once used by HERV-T.

Chapter VI. Origin and evolution of an orphan antiviral gene

The other side of paleovirology is the study of the host antiviral defenses and the effects that ancient viruses have had on them. As described in chapter I, the *tetherin* gene encodes an antiviral protein with no known homologs that traps a broad spectrum of enveloped viruses at infected cell surfaces. Its highly unusual architecture (TM-CC-GPI, chapter I), rather than primary sequence, is critical for Tetherin function (Perez-Caballero et al., 2009). The uniqueness of this antiviral mechanism, along with the absence of any known homolog gene or analogous process in the normal functioning of cells prompted us to ask the question of how this gene and antiviral mechanism arose. The body of work presented in this chapter was produced in collaboration with Dr. Siddarth Venkatesh.

Candidate ancestors of tetherin.

Every sequenced genome harbors a significant fraction of "orphan" genes whose origins are obscure due to the absence of homologs (Khalturin et al., 2009; Palmieri et al., 2014; Tautz and Domazet-Loso, 2011). Unlike other antiviral restriction factors, *tetherin* is an orphan gene in eutherian mammals whose origins are unknown. Moreover, *tetherin* is a non-essential gene in mice (Liberatore and Bieniasz, 2011) and is only expressed in response to interferon in most cells. Thus, it is difficult to envisage a scenario wherein *tetherin* arose from a gene of similar, necessary function. Homology searches for possible ancestors of *tetherin* have been unsuccessful. However, orphan genes that are

unusually divergent might preserve structural but not sequence similarity with their parents (Domazet-Loso and Tautz, 2003). Therefore, we searched for candidate sister genes that encode proteins of similar predicted architecture, irrespective of its sequence. We initially focused on the human and mouse genomes because they are well annotated and regularly updated. We searched for annotated genes whose products are predicted to have the following features: (i) a single TM domain, (ii) a CC domain C-terminal to the TM domain, and, (iii) a GPI modified C-terminus (see chapter II). Among the 22,691 and 22,709 annotated human and mouse protein coding genes (Ensembl Release 71) (Flicek et al., 2013), Tetherin was the only protein that is predicted to have a TM-CC-GPI architecture.

We reasoned that the genesis of *tetherin* might have involved the acquisition of a third domain by gene products encoding 2 out of the 3 contiguous Tetherin features, i.e. proteins with either TM-CC or CC-GPI domains. A search of the Ensembl database found 66 to 211 human and 47 to 175 mouse TM-CC proteins, depending on the length requirement imposed to form a bona fide coiled coil (Table 6.1). Analysis of a previously predicted set of GPI-modified human proteins revealed that Tetherin and 12 other proteins had a CC-GPI configuration (Pierleoni et al., 2008) (Table 6.2). Overall, ~1% of all annotated human or mouse genes encode proteins that have TM-CC or CC-GPI architectures from which Tetherin could plausibly have arisen through acquisition of its third defining feature.

Table 6.1: Identification of genes in the human and mouse genomes that encode TM-CC proteins

Species	Minimum Length of Coiled-coil domain	TM-CC proteins	Homologous adjacent gene pairs (same orientation)	Non-homologous adjacent gene pairs (same orientation)	Non-homologous adjacent gene pairs (reverse orientation)
Human	**21**	211	**7:** AREG and AREGB, PCDHA4 and PCDHA5, PCDHA8 and PCDHA9, PCDHA9 and PCDHA10, PCDHA11 and PCDHA12, NDUFA13 and YJEFN3, FCER2 and CLEC4G	**3:** BCAP29 and SLC26A4, CLEC4M and EVI5L, PV1 and Tetherin	**1:** EVC and EVC2
	28	107	**1:** FCER2 and CLEC4G	**1:** PV1 and Tetherin	**1:** EVC and EVC2
	35	78	**1:** FCER2 and CLEC4G	**1:** PV1 and Tetherin	**1:** EVC and EVC2
	42	66	**0**	**1:** PV1 and Tetherin	**1:** EVC and EVC2

Continuation of Table 6.1

Species	Minimum Length of Coiled-coil domain	TM-CC proteins	Homologous adjacent gene pairs (same orientation)	Non-homologous adjacent gene pairs (same orientation)	Non-homologous adjacent gene pairs (reverse orientation)
Mouse	**21**	175	**5:** CLEC4F and CD207, CYP2B9 and CYP2B13, BC096441 and TNFSF12, MGL2 and CLEC10A, FCER2 and CLEC4G	**1:** PV1 and Tetherin	**3:** BC035947 and MOGAT1, EVC and EVC2, CYP4B1 and EFCAB14
	28	81	**3:** CLEC4F and CD207, MGL2 and CLEC10A, FCER2 and CLEC4G	**1:** PV1 and Tetherin	**1:** EVC and EVC2
	35	57	**3:** CLEC4F and CD207, MGL2 and CLEC10A, FCER2 and CLEC4G	**1:** PV1 and Tetherin	**1:** EVC and EVC2
	42	47	**1:** CLEC4F and CD207	**0**	**1:** EVC and EVC2

*Number of homologous and non-homologous gene pairs depending on the threshold set for the length of the coiled-coil domain.

Table 6.2: Identification of genes in the human genomes that encode CC-GPI proteins

Associated Gene Name	Chromosome	Gene Start (bp)	Gene End (bp)	Strand[1]	Length (aa)	TM number[2]	CC blocks[3]	Length of the CC domain (aa)[4]	Omega site[5]
MIER1	1	67153227	67226890	1	451	0	1	21	419
BCAN	1	154878357	154895939	1	672	0	1	35	646
TMEM165	4	55956877	55987095	1	325	7	1	24	296
VNN2	6	133106703	133126291	-1	521	0	1	21	491
ULBP3	6	150425979	150431924	-1	245	0	1	21	222
CCDC136	7	128218784	128249419	1	435	0	5	253	403
CACNA2D4	12	1771384	1898131	-1	1138	0	1	22	1109
SLC39A2	14	20537259	20539870	1	310	8	1	21	284
OTOA	16	21597336	21679551	1	816	0	1	21	790
Tetherin	19	17374750	17377401	-1	181	1	1	49	157
SLC8A2	19	52623735	52666934	-1	826	3	1	21	797
EGFL6	X	13497645	13561614	1	555	0	1	28	529

[1] Orientation of the gene. [2] Number of TM domains predicted by TMHMM. [3] Number of CC blocks predicted by COILS. [4] Number of amino acids in the coiled-coil domains. [5] Omega site predicted using Pred-GPI.

One likely mechanism by which *tetherin* arose is the duplication and neofunctionalization of another gene (Ohno, 1970). Because duplicated genes are often positioned adjacent to each other in genomes (Pan and Zhang, 2008), we inspected the organization of genes proximal to *tetherin* in human and mouse genomes. Strikingly, two genes whose protein product had a TM-CC configuration were found to be positioned proximal to *tetherin* in human and mouse genomes (Figure 6.1A). One such gene is *pv1* that encodes an essential component of the stomatal diaphragms of caveolae and the diaphragms of fenestrae and transendothelial channels (Stan et al., 1999b; Stan et al., 2012). Another TM-CC gene, that we have designated TM-CC, adjacent to *tetherin*, *tm-cc(at)*, is of unknown function and is located between *pv1* and *tetherin* (Figure 6.1A). *tm-cc(at)* was not identified in our initial searches of mouse and human genomes because it is not annotated in the Ensembl database and is the translated product of a Gnomon-predicted transcript (GenBank: XM_011242347.1) whose existence in mammals is supported only by a single cDNA sequence (GenBank: AK141396.1, RIKEN C430049E01) from an early mouse embryo. Thus, it was initially uncertain whether tm-cc(at) is a bona fide gene. However, PCR analysis of day 7 mouse embryo mRNA confirmed that a transcript encoding *tm-cc(at)* was present therein, and spliced at the predicted intron-exon junctions (Figure 6.2A). Additionally, *tm-cc(at)* exons appear to be better conserved compared to flanking genome sequence across mammals (Figure 6.2B), again supporting its existence as a bona fide gene. Moreover, a

cDNA encoding a homologous protein has been found in the chicken transcriptome (GenBank: BU332041.1, cDNA clone ChEST413k13). Even though there is no mRNA evidence for *tm-cc(at)* expression in humans, a homologous sequence is present in the syntenic genomic location of humans (GenBank: XM_011528476.1) and RNA-seq data suggest that it is expressed at low abundance and is spliced in a similar way to mouse TM-CC(aT), but includes a fifth exon (Figure 6.2C). Thus, this sequence appears to be (or has been) a gene that is poorly annotated because it is rarely expressed.

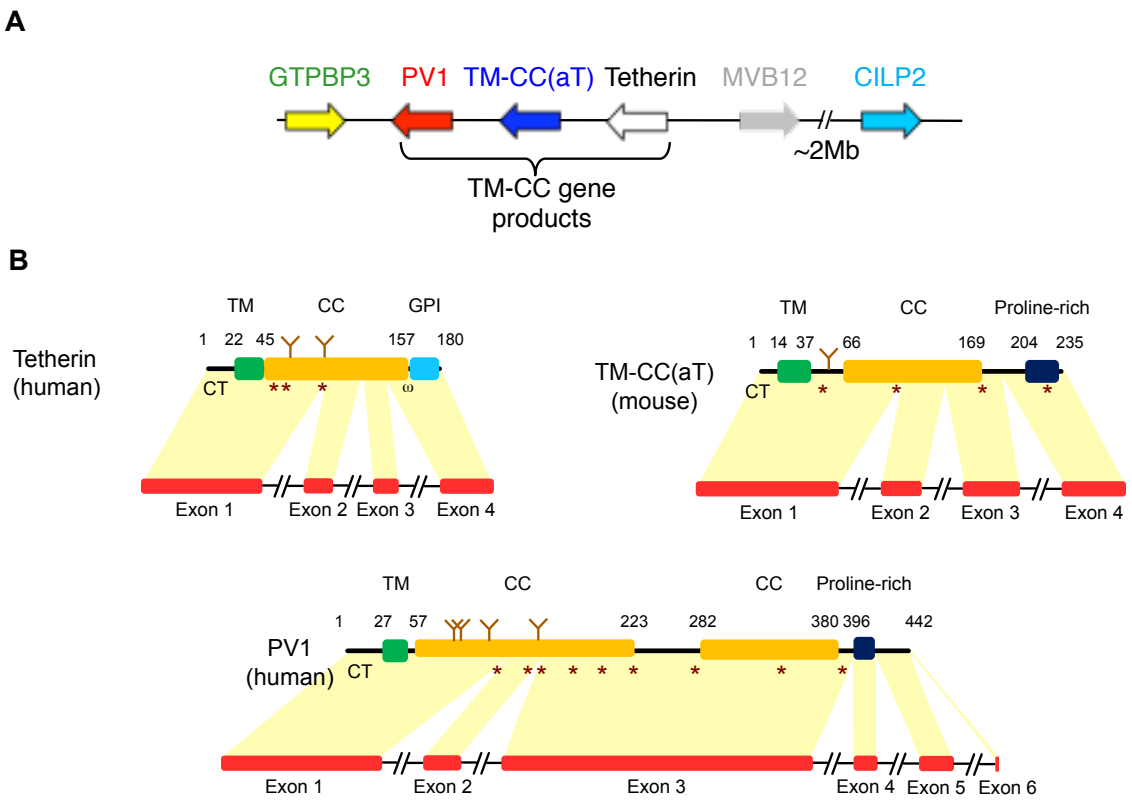


Figure 6.1: Location and architecture of TM-CC gene products proximal to *tetherin* in human and mouse genomes.

(A) Diagram of genes surrounding *tetherin* in the human and mouse genomes, generated using NCBI and UCSC Genome Browsers. **(B)** Organization of TM-CC genes and their protein products for human Tetherin (GenBank: NP_004326.1), mouse TM-CC(aT) (GenBank: XP_003945491.1), and human PV1 (GenBank: NP_112600.1) proteins. Glycosylation and cysteine residues are indicated as brown Y symbols and stars respectively. Numbers indicate amino acid positions. Structural features of TM-CC(aT) and PV1 are based on predictions using TMHMM (Krogh et al., 2001), COILS (Lupas et al., 1991), and Pred-GPI (Pierleoni et al., 2008). Tetherin features are based on structural and functional data.

Figure 6.2: Validation of *tm-cc(at)* as a bona-fide gene.

(A) PCR amplification of *tm-cc(at)* from mouse total RNA using (i) a forward primer specific to the first exon and a reverse primer specific to the third exon, and (ii) a forward primer specific to the third exon and a reverse primer specific to the 3'UTR. L: DNA ladder. Lanes (i) and (ii). PCR amplicons from cDNA derived from 7-day mouse embryo RNA. The transcript structure of mouse *tm-cc(at)* and the expected 280 bp and 867 bp amplicons are indicated. **(B)** Evolution of *tm-cc(at)*. Graphical representation of a glocal sequence alignment of human *tm-cc(at)* with its orthologs in the indicated animal species generated using mVISTA (Frazer et al., 2004). Each graph shows the percentage of conservation between the DNA sequence of that organism with the corresponding human sequence at any given coordinate. Introns that cross the 50% sequence similarity threshold are indicated as pink regions, whereas coding exons are indicated as purple regions and untranslated regions are indicated as light blue regions. **(C)** Alignment of human (adapted from GenBank: XP_011526778.1), mouse (GenBank: XP_003945491.1) and chicken (adapted from GenBank: XP_015155597.1) TM-CC(aT) protein sequences. Sequences spanning the cytoplasmic tail (CT), transmembrane domain (TM) and extracellular coiled-coil domain (CC) are indicated. Residues that comprise the proline-rich domain are indicated in red.

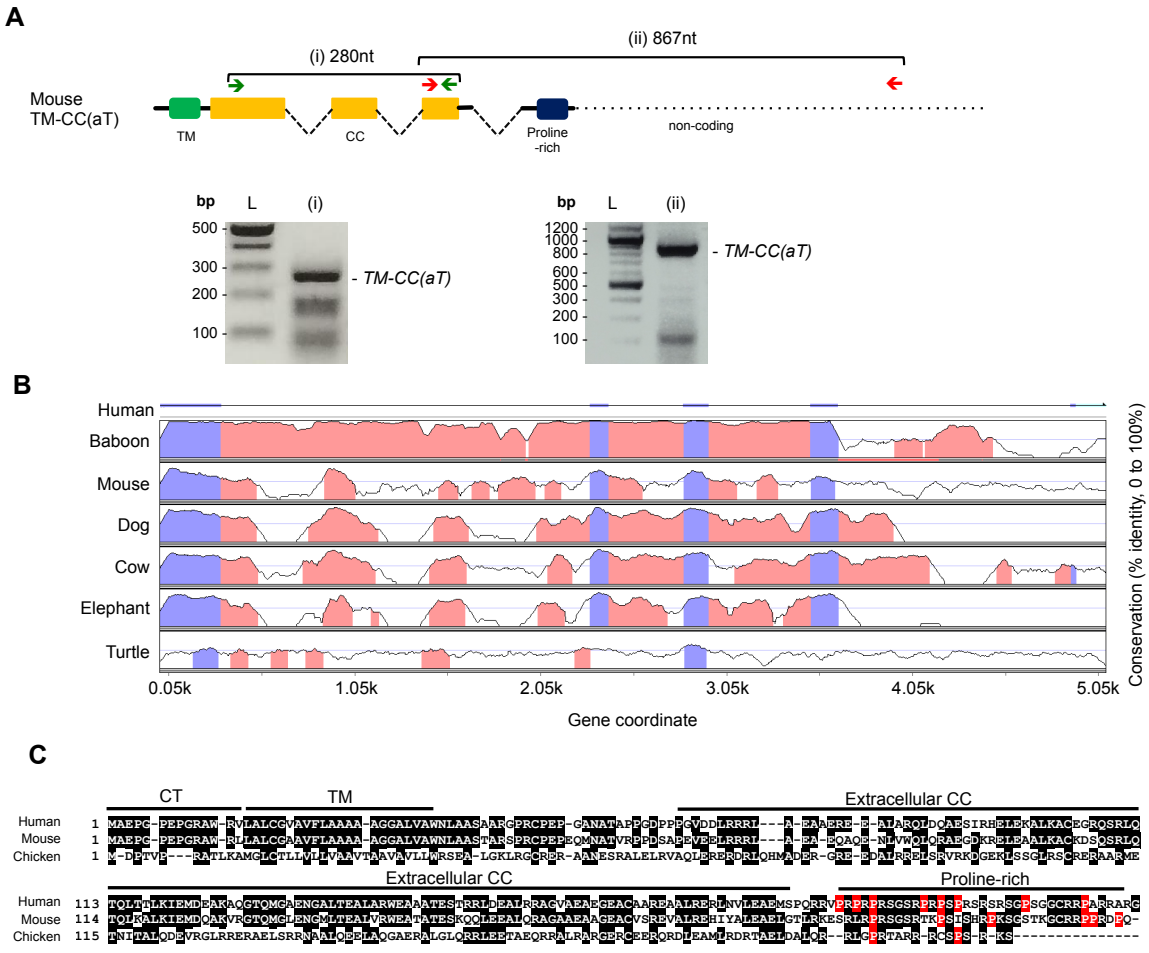


Figure 6.2: Validation of *tm-cc(at)* as a bona-fide gene.

Tetherin, PV1 and TM-CC(aT) proteins have a short N-terminal CT, a single TM domain, and a predominantly CC extracellular domain with multiple cysteine residues that, in the case of Tetherin and PV1, are known to stabilize parallel homodimer formation via the formation of disulfide bonds (Andrew et al., 2009; Perez-Caballero et al., 2009; Stan et al., 1999a) (Figure 6.1B). Analysis of intron-exon architecture reveals that *tetherin* has 4 exons, *pv1* has 5-7 exons (depending on the species) and cDNA sequences indicate that *tm-cc(at)* has 4 (mouse) or 5 (chicken) exons. The organization of exons and protein coding domains in *tetherin*, *pv1* and *tm-cc(at)* is similar, particularly when *tetherin* and *tm-cc(at)* are compared, with the first exon of all three genes encoding the CT, TM and the N-terminal portion of their CC domains. Notably, human and mouse *pv1* and *tm-cc(at)* both encode proline-rich sequences at their C-termini, rather than the GPI anchor encoded by Tetherin (Figure 6.1B). The unusual structural similarity of these genes and proteins, along with their adjacent genomic location, suggests that *pv1*, *tm-cc(at)* and *tetherin* might share a common ancestor, despite the absence of sequence similarity.

To determine how unexpected this apparent clustering of TM-CC encoding genes should be, we looked at the annotated human and mouse genes for adjacently positioned pairs that encode TM-CC proteins. Among the 211 and 175 TM-CC-encoding genes in the human and mouse genomes respectively, 10 (human) and 6 (mouse) gene pairs were found adjacent to each other and are in the same orientation (Table 6.1). Seven of 10 TM-CC gene pairs in humans, and 5 of 6

pairs in mice, share clear sequence similarity usually belonging to obvious gene families (e.g. CLEG and CTEG proteins) and therefore very likely arose via gene duplication (Table 6.1), in accordance to the fact that most adjacent duplicated genes (72-94%) occur in the same orientation (Pan and Zhang, 2008). Overall, TM-CC genes account for ~1% of all annotated genes, and most of the few adjacent TM-CC gene pairs obviously arose via duplication of the neighboring gene. Thus, the probability that *tetherin* originated de novo in a distal chromosomal location, and was then inserted adjacent to two genes that encode proteins of similar TM-CC architecture appears to be low (< ~211 TM-CC encoding genes / 22,691 genes in the human genome, or < 0.01).

TM-CC(aT) and PV1 can be endowed with antiviral activity by the addition of a GPI anchor.

Both the TM and GPI anchor domains present in Tetherin are essential for virion entrapment at the cell surface (Perez-Caballero et al., 2009). A clear difference in the overall architectures of human and mouse PV1, TM-CC(aT) and Tetherin is the presence of the GPI anchor in Tetherin. If the three proteins indeed share a common ancestor, then a model for the genesis of Tetherin would include acquisition of a GPI anchor by an ancestral duplicated gene. Indeed, given the apparent plasticity of Tetherin protein sequences, this simple modification might be sufficient to endow PV1 and/or TM-CC(aT) with antiviral activity.

We measured the yield of HIV-1 particles from transfected 293T cells expressing unmodified PV1 and TM-CC(aT) proteins, or derivatives of PV-1 and TM-CC(aT) proteins that were appended with a GPI modification signal from human Tetherin. This analysis revealed that the engineered human and mouse TM-CC(aT) proteins with GPI modified C-termini inhibited HIV-1 particle release nearly as potently as human Tetherin (Figures 6.3A and B). Interestingly, the unmodified TM-CC(aT) proteins also had some propensity to trap virions, while the unmodified PV1 protein had little if any antiviral activity (Figures 6.3A and B). The addition of a GPI anchor provided PV1 with virion entrapment activity that was less potent than that exhibited by the GPI-modified TM-CC(aT) proteins (Figures 6.3A and B). Western blot analysis of N-terminally tagged derivatives of these proteins confirmed that active and inactive proteins were approximately equivalently expressed (Figure 6.3C).

We assessed the antiviral activity of two unrelated TM-CC proteins, CD72 and CLEC1A, which have been demonstrated to form dimers and to reside at the cell surface (Sattler et al., 2012; Von Hoegen et al., 1990). The unmodified CD72 protein had a minor propensity to inhibit HIV-1 virion release (Figure 6.4). However, unlike PV1 or TMCC(aT), the addition of a GPI anchor did not confer either protein with the ability to trap virions (Figure 6.4), suggesting that the ability to inhibit virion release upon acquisition of a C-terminal GPI anchor is not generalizable to all TM-CC proteins.

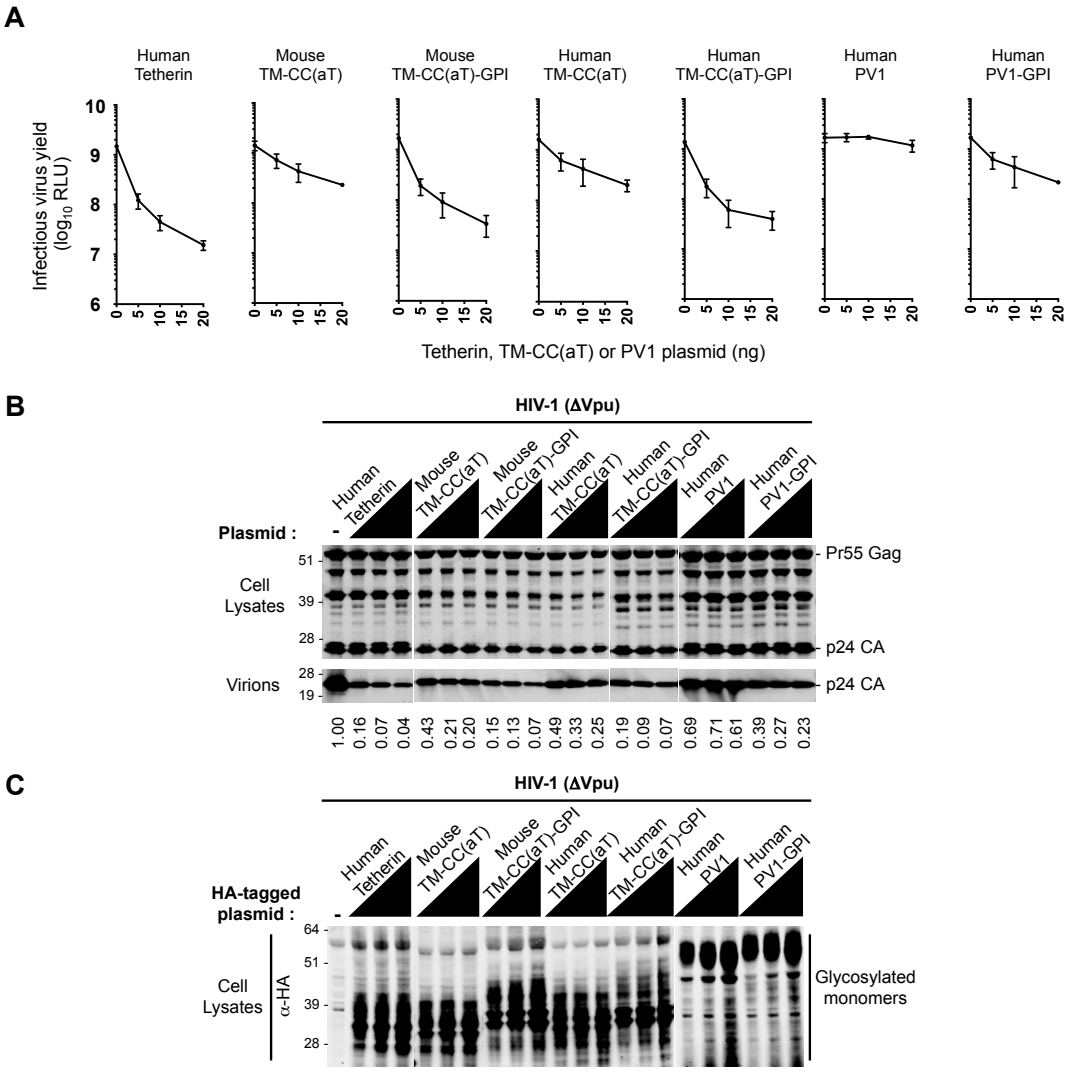


Figure 6.3: Antiviral activity of GPI-modified TM-CC(aT) and PV1 proteins.

(A) Infectious virion yield measured using HeLa TZM-bl indicator cells following transfection with a Vpu-deficient HIV-1 proviral plasmid along with increasing amounts of the indicated unmodified or GPI-modified Tetherin, TM-CC(aT) or PV1 proteins. RLU: relative light units. Data from 3 independent experiments. **(B)** Western blot analyses (anti-CA) of cell lysates and virions corresponding to (A). Numbers at the bottom represent virion CA protein levels relative to those obtained in the absence of an inhibitor. **(C)** Western blot analyses following transfection of plasmids expressing HA-tagged Tetherin or GPI-modified TM-CC(aT) and PV1 proteins. The cell lysates were probed with anti-HA antibodies.

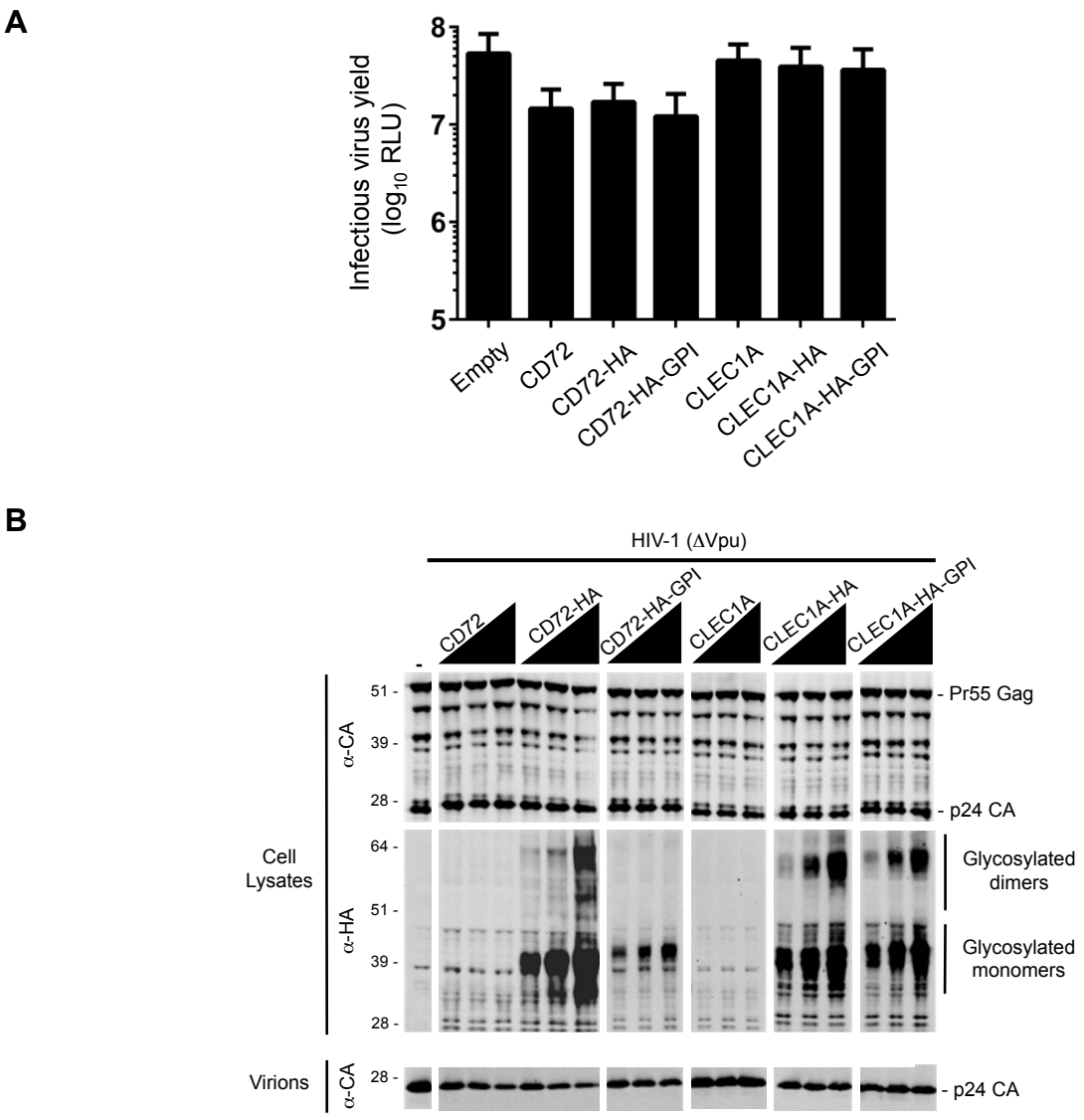


Figure 6.4: Antiviral activity of unrelated TM-CC proteins.

(A) Infectious virion yield measured using HeLa TZM-bl indicator cells following transfection with Vpu-deficient proviral plasmid along with plasmids expressing the modified CD72 and CLEC1A proteins. RLU: relative light units. Data from 3 independent experiments. (B) Western blot analyses of transfected 293T cell lysates and virions corresponding to (A), except that varying amounts of the indicated plasmids were used. The cell lysates were probed with anti-CA and anti-HA antibodies. Virions were probed with anti-CA antibody.

Sequence and structural homologs of Tetherin in diverse vertebrates.

Typically, sister genes exhibit sequence similarity, but homology was not evident in human or mouse PV1, Tetherin and TM-CC(aT) proteins or genes. Therefore, we next sought to delineate the evolutionary history of *tetherin* by identifying homologs in diverse species. BLAST searches revealed that homologs of human Tetherin are present in therian mammals (marsupials and eutherian mammals) (Figure 6.5) suggesting an origin predating the emergence of mammals. Tetherin sequence homologs could not be found in any other vertebrate species by BLAST. However, the marginal sequence similarity among some mammalian Tetherin proteins suggested that sequence divergence over >150 million years may have eroded sequence similarity to a point that orthologs might not be detected in more diverse species, using only their sequences.

Therefore, we next collected all annotated and predicted transcript sequences (Gnomon (Souvorov et al., 2010)) from representative diverse species of monotremes (platypus), birds (chicken, saker falcon, turkey), reptiles (Chinese alligator and painted turtle), amphibians (frog), lobe-finned fish (coelacanth), ray-finned fish (medaka and zebrafish) and cartilaginous fish (elephant shark). We also obtained transcriptomic data for a jawless vertebrate, the sea lamprey (Smith et al., 2013). We searched these actual and putative transcripts for sequences that were predicted to encode TM-CC-GPI proteins. By this approach we were able to identify putative Tetherin-like proteins (i.e. proteins exhibiting a TM-CC-GPI architecture, but lacking sequence homology to Tetherin) in most of

the aforementioned species (Figure 6.5) except platypus, chicken, turkey and frog. There was no statistically significant sequence similarity between these TM-CC-GPI proteins and known or putative Tetherin proteins from mammals. Thus, if the genes encoding these proteins share a common ancestor with Tetherin, they diverged from all previously analyzed eutherian mammal proteins ~150 to 500 MYA (Inoue et al., 2010; Janvier, 2006; Venkatesh et al., 2014).

Although an inspection of other bird genomes did not yield any Tetherin homologs, we found homologs of the saker falcon TM-CC-GPI protein in eagles (Canadian eagle and the bald eagle), ibises (crested ibis), penguins (emperor penguin), hummingbirds (Anna's hummingbird), cuckoos (common cuckoo) and other falcons (peregrine falcon); some of these are supported by transcriptomic evidence (Canadian eagle, bald eagle, emperor penguin and saker falcon). BLAST searches also revealed sequences in the turkey genome that were similar to the falcon TM-CC-GPI protein in the predicted terminal four exons of a proximal gene (*cilp2*) (Figure 6.5). However, the inclusion of these four exons in turkey cilp2 is unsupported by RNA-seq data and likely represents an annotation error. We also found sequences adjacent to the chicken *tm-cc(at)* gene that are predicted to encode a TM-CC-GPI protein. These observations suggest that the genomes of both Neoaves (i. e. falcons, eagles, ibises, etc.) and Galloanseres (i.e. chicken and turkey) have the potential to code for Tetherin/TM-CC-GPI proteins.

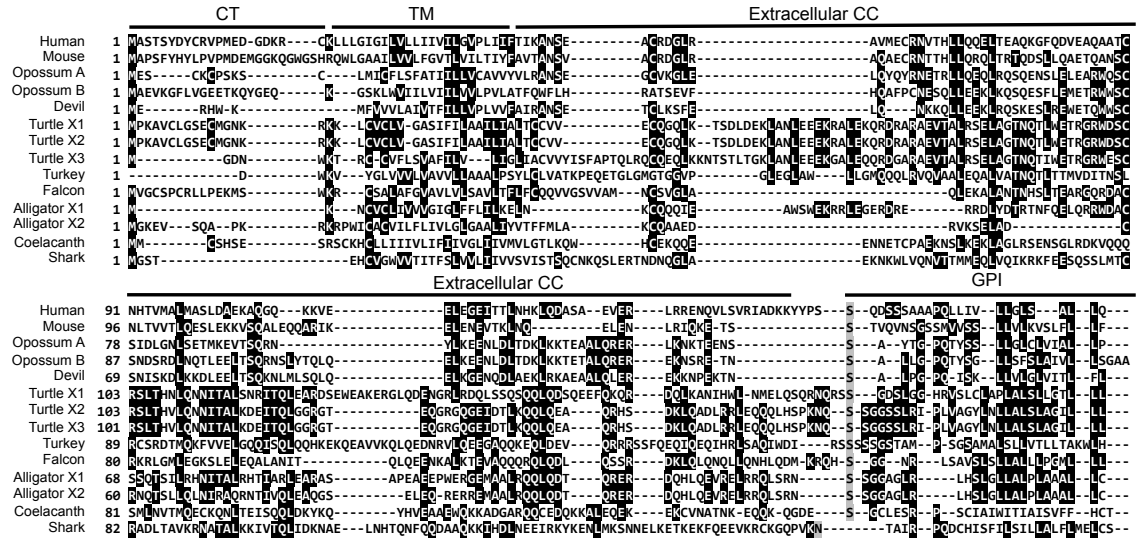


Figure 6.5: Alignment of Tetherin/TM-CC-GPI protein sequences.

Human (GenBank: NP_004326.1), mouse (GenBank: NP_932763.1), opossum (GenBank: XP_007489270.1 and XP_007489271.1), Tasmanian devil (GenBank: XP_012399618.1), turtle (GenBank: XP_008169758.1, XP_005279001.1 and XP_005279003.1), turkey (inferred from GenBank: XP_010723307.1), falcon (inferred from GenBank: XP_005444407.1 and Gnomon prediction: 2189215010.p), alligator (GenBank: XP_006017475.1 and XP_006017476.1), coelacanth (Gnomon prediction: 16424589.p), elephant shark (GenBank: XP_007897024.1). Sequences spanning the TM, CC domains and GPI anchor are indicated. Conserved residues are highlighted and predicted omega (GPI modification) sites are indicated in grey.

Antiviral activity of divergent Tetherin and TM-CC-GPI proteins.

Previously, only Tetherin proteins encoded by eutherian mammals and by one reptile have been identified and demonstrated to exhibit strong antiviral activity (Arnaud et al., 2010; Heusinger et al., 2015; Takeda et al., 2012). We next determined whether widely divergent Tetherin homologs encoded by marsupials (opossum and Tasmanian devil), as well as the TM-CC-GPI proteins encoded by a bird (saker falcon), a reptile (Chinese alligator) a lobe-finned fish (coelacanth), a cartilaginous fish (elephant shark), and a jawless fish (lamprey), possessed antiviral activity.

We transfected 293T cells with a panel of plasmids encoding Tetherin/TM-CC-GPI proteins, along with a Vpu-deficient HIV-1 (ΔVpu) proviral plasmid. Because antibodies to these proteins were not available, we tested authentic untagged proteins as well as derivatives encoding HA-epitope tags at their amino termini in most cases (Figure 6.6). The putative Tetherin/TM-CC-GPI proteins encoded by opossum, Tasmanian devil, alligator, falcon, coelacanth and elephant shark, all inhibited HIV-1 virion release, in most cases with an apparent potency that was similar to human Tetherin (Figure 6.6). The addition of a N-terminal HA tag verified expression (Figure 6.6C), but affected potency in some cases (not shown). In contrast, the TM-CC-GPI protein encoded in the lamprey genome was poorly expressed and inactive (Figures 6.6A, B and C). Two divergent Tetherin proteins (Tasmanian devil and Chinese alligator) were tested for susceptibility to antagonism by HIV-1 Vpu and, predictably, were found to be resistant (Figure

6.6D and E). These findings suggest that virion entrapment is a nearly universal feature of TM-CC-GPI proteins that have arisen in jawed vertebrates within the past ~450 million years.

Genomic loci harboring tetherin/TM-CC-GPI genes.

In eutherian mammals, *tetherin* and its neighboring genes form a syntenic block: *gtpbp3–pv1–tm-cc(at)–tetherin–mvb12*–[2Mb]–*cilp2*) (Figure 6.1A). While most mammals possess a single *tetherin* gene, recent duplications have led to the presence of multiple homologous *tetherin* genes in some species, such as cows, sheep, and bats (Arnaud et al., 2010; Takeda et al., 2012). Inspection of this genomic locus in diverse species revealed that in the opossum and the wallaby, *tetherin* has been recently duplicated, *tm-cc(at)* is absent, and *pv1* is separated from the two *tetherin* genes by ~31 mega bases (MB). Moreover, the *tetherin* genes are adjacent to a gene, *cilp2* that is located ~2 MB distal to *tetherin* in eutherian mammals (Figure 6.1A and 6.7A). We could not reconstruct an analogous locus in a monotreme (platypus) because its genome is incompletely assembled.

Remarkably, we found that all of the aforementioned TM-CC-GPI encoding genes in avians, reptiles, coelacanth, medaka and shark species were present in nearly the same location in their respective genomes, i.e. between *pv1* and *cilp2* (Figure 6.7A). By manual curation of genomic sequence in this interval in some species, including searches of translated genomic sequences, and inspection of RNAseq

data, we also found that there likely have been several duplication and deletion events involving *tm-cc(at)* and/or *tm-cc-gpi* genes in the *gtpbp3–pv1–cilp2* interval during the course of vertebrate evolution (Figure 6.7). For example, single *tm-cc-gpi* and *tm-cc(at)* genes are present in most eutherian mammals and avian species, but in reptiles, there is a single *tm-cc(at)* gene and multiple *tm-cc-gpi* genes including some that appear to generate different TM-CC-GPI protein isoforms via alternative splicing of duplicated exons (Figure 6.7B). In an amphibian (frog) the locus lacks a *tetherin* homolog or a putative *tetherin* gene. In the coelacanth, a single *tm-cc-gpi* and three *tm-cc(at)* genes are present. In a ray-finned fish (medaka), a *tm-cc-gpi* gene is linked to this locus, but positioned outside the *gtpbp3–cilp2* interval and the genes are arranged in a manner consistent with the possibility that a segmental inversion has occurred. In the elephant shark genome, *tm-cc(at)* is absent and a *tm-cc-gpi* gene is present in this interval proximal to *cilp2* (separated from it by an intron-less gene) and separated from *pv1* by ~223 KB (Figure 6.7A). It was not possible to reconstruct the configuration of this locus in the lamprey due to the fragmented nature of the genome sequence. Nevertheless, these findings suggest that one or more *tm-cc(at)* and/or *tm-cc-gpi* genes, including *tetherin*, appeared in vertebrates in at the genomic locus containing *gtpbp3–pv1–cilp2* (Figure 6.7).

Figure 6.6: Antiviral activity of divergent Tetherin/TM-CC-GPI proteins.

(A) Infectious virion yield measured using HeLa TZM-bl indicator cells following transfection of Vpu-deficient HIV-1 proviral plasmids along with plasmids expressing Tetherin/TM-CC-GPI proteins. RLU: relative light units. Data from 3 independent experiments. **(B)** Western blot analyses (anti-CA) of cell lysates and virions corresponding to (A). Numbers at the bottom represent virion CA protein levels relative to those obtained in the absence of an inhibitor. **(C)** Western blot analyses following transfection of Vpu-deficient HIV-1 proviral plasmids along with plasmids expressing HA-tagged Tetherin/TM-CC-GPI proteins. The cell lysates were probed with anti-HA antibodies. **(D)** Infectious virion yield measured using HeLa TZM-bl indicator cells and given in relative light units (RLU) following transfection of wild type (WT) HIV-1 proviral plasmids along with plasmids expressing Tetherin proteins in a eutherian mammal (human), a marsupial (Tasmanian devil) and a reptile (Chinese alligator). **(E)** Western blot analyses (anti-CA) of cell lysates and virions corresponding to (D). Numbers at the bottom represent virion CA protein levels relative to those obtained in the absence of an inhibitor.

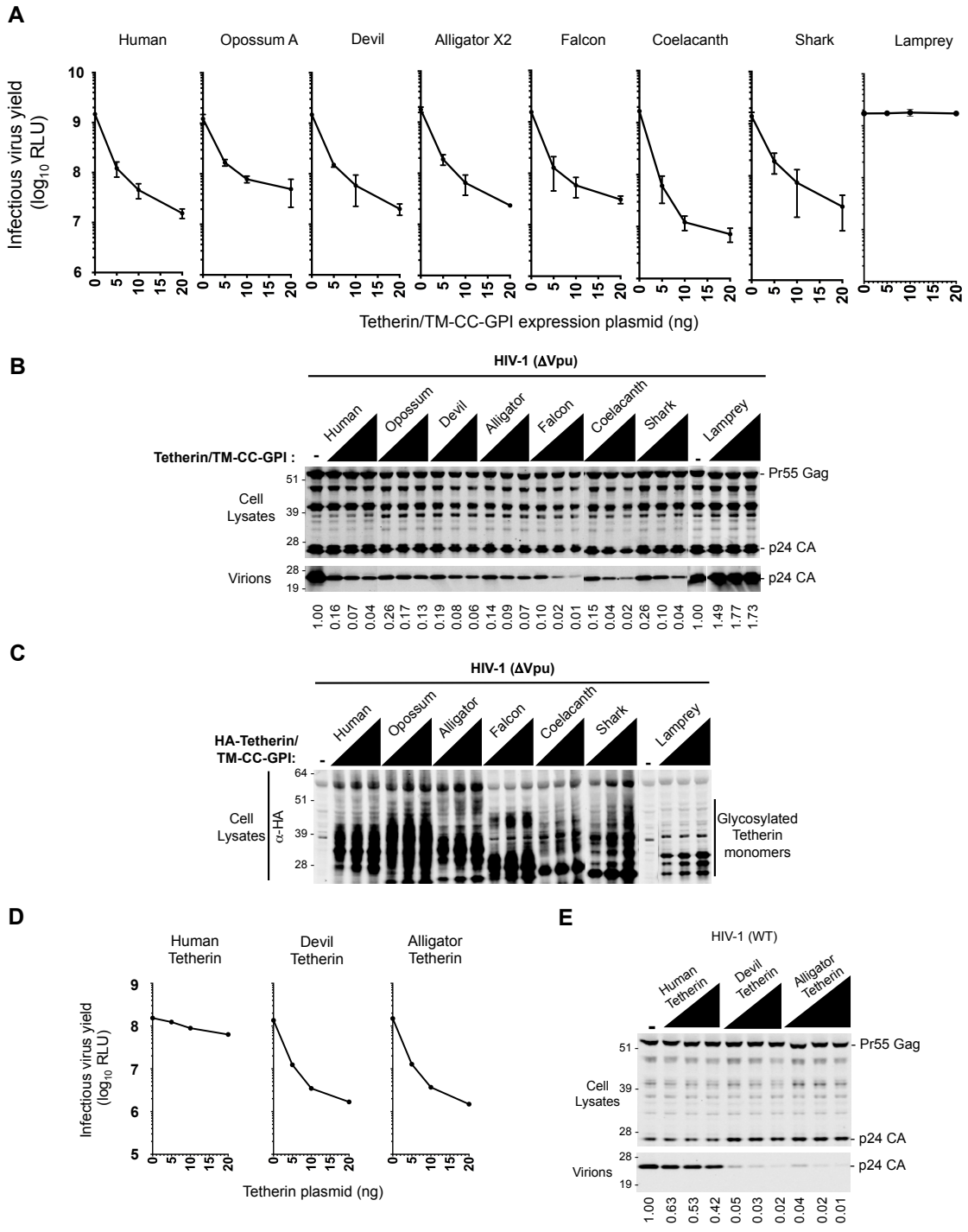


Figure 6.6: Antiviral activity of divergent Tetherin/TM-CC-GPI proteins.

Figure 6.7: Organization of genes in the *tetherin* locus.

(A) Diagrams were generated using NCBI, UCSC, Ensembl Genome Browsers and sequence similarity approaches. Branches in gray indicate incomplete genome assemblies. Proteins with TM-CC and TM-CC-GPI structures were identified in annotated, predicted proteins or RNA-seq data using TMHMM (Krogh et al., 2001), COILS (Lupas et al., 1991), and Pred-GPI (Pierleoni et al., 2008). Inclined figures indicate genes in incompletely assembled scaffolds. White and black stars indicate genes that were active or not active respectively in virion release-inhibition assays. Potential alternatively spliced versions of alligator and turtle *tetherin* are indicated by dotted lines. Phylogeny and speciation dates were based on (Inoue et al., 2010; Janvier, 2006; Venkatesh et al., 2014). **(B)** Genomic loci and spliced isoforms of human, alligator and turtle *tetherin/tm-cc-gpi*. CT, TM, CC and GPI anchor are indicated in color. GenBank accession numbers are indicated in brackets. Gray shaded forms indicate the relationship between a specific genomic position and the transcript variant it encodes.

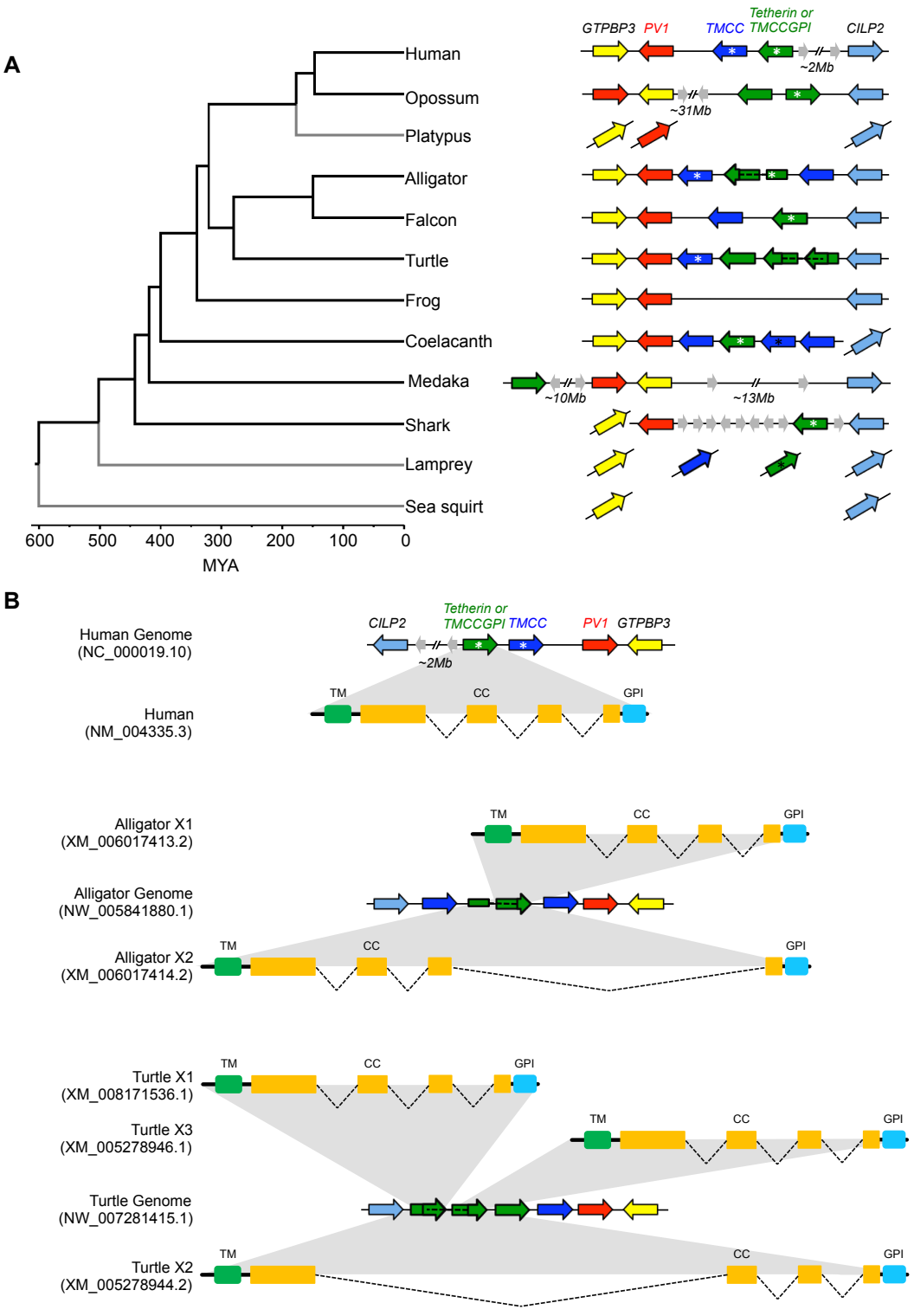


Figure 6.7: Organization of genes in the *tetherin* locus.

Apparent differences in selective pressures acting on pv1, tm-cc(at) and tetherin

pv1, *tm-cc(at)* and *tetherin* do not exhibit sequence similarity to each other in mammals. If the hypothesis that they indeed share a common ancestor is correct, then a key question is whether they should be expected to share sequence similarity, given the time at which they diverged, and the types of selection pressure that have acted on them.

We first determined the types of selection pressures to which *pv1*, *tm-cc(at)* and *tetherin* have been subjected, using codon-based tests in CODEML (Yang, 1997) and nested pairs of models (M0 and M3, M1a and M2a, M7 and M8) (Table 6.3). Because sequence divergence in other animal species confounded the reliable assignment of homologous sites in Tetherin, (Figure 6.5) we restricted our analyses to therian mammals, which have diverged over the past ~150 million years, Positive selection appears to have influenced *tetherin* evolution in most therian mammals (M1a vs. M2a: p-value < 0.002; M7 vs. M8: p-value < 0.001) (Table 6.3). In contrast, *pv1* and *tm-cc(at)* have evolved under purifying selection in therian mammals (dN/dS < 1). Although there was evidence for variable selective pressures among *pv1* and *tm-cc(at)* sites (M0 vs M3: p-value < 0.001) (Table 6.3), this variation does not appear to be explained by positive selection (M1a vs. M2a: p-value > 0.99; M7 vs. M8: p-value > 0.01; no positively selected sites with a PP > 0.95) (Table 6.3). Overall, these results are consistent with the notion that *pv1*, and perhaps *tm-cc(at)*, have conserved cellular functions in

therian mammals, while varying pressures in different mammalian species have selected for changes in *tetherin* sequences.

Notably, introns share little sequence similarity across *pv1* in divergent amniotes, indicating that neutral evolution over ~300 MY is sufficient to diminish sequence similarity to nearly undetectable levels at presumed neutrally evolving sites (Figure 6.8). Thus, genetic drift followed by the very different selection pressures imposed on *pv1*, *tm-cc(at)* and *tetherin*, can explain the paucity of sequence similarity among modern *tm-cc* genes that putatively originated from a single common ancestor early in vertebrate evolution.

Table 6.3: Likelihood ratio test on nested models of variable ω ratios among sites and Naive Empirical Bayes (NEB) probabilities per site for TM-CC(aT), PV1 and Tetherin coding sequences.

Gene	Data Set	-2 ln(λ)	df	P-value	PSS
Tetherin	M0 vs. M3	152.675032	4	P < 0.001	4
	M1 vs. M2	13.171016	2	0.0014	3
	M7 vs. M8	19.830186	2	P < 0.001	4
PV1	M0 vs. M3	852.432408	4	P < 0.001	0
	M1 vs. M2	0.000366	2	0.9998 (NS)	0
	M7 vs. M8	8.029462	2	0.0180 (NS)	0
TM-CC(aT)	M0 vs. M3	198.13646	4	P < 0.001	0
	M1 vs. M2	-2E-06	2	NA (NS)	0
	M7 vs. M8	1.160924	2	0.5596 (NS)	0

Neutral models of selection (M0, M1, M7) were compared to models that allow variation in dN/dS among sites (M3) or selection models (M2, M8). p-values were calculated using a chi square distribution. df: degrees of freedom. PSS: number of positive selection sites (Naïve Empirical Bayes) with posterior probabilities > 0.95. NS: Non-significant. NA: p-value ~ 1.

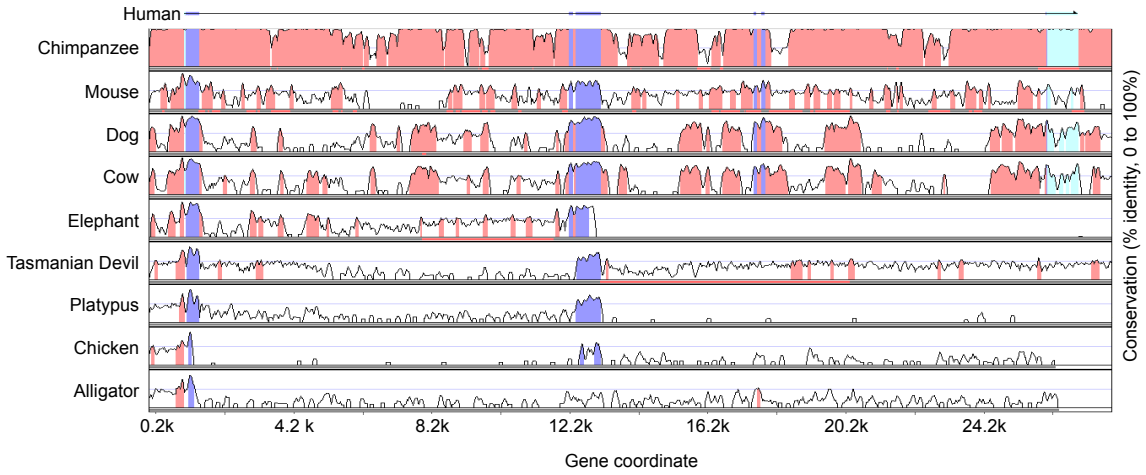


Figure 6.8: Sequence conservation of the *pv1* gene.

Graphical representation of a glocal sequence alignment of human *pv1* with its orthologs in the indicated animal species generated using mVISTA (Frazer et al., 2004). Each graph shows the percentage of conservation between the DNA sequence of that organism with the corresponding human sequence at any given coordinate. Introns that cross the 50% sequence similarity threshold are indicated as pink regions, whereas coding exons are indicated as purple regions and untranslated regions are indicated as light blue regions.

Relationship between PV1, TM-CC(aT) and tetherin/TM-CC-GPI proteins and genes.

Although these findings suggest that tetherin arose from duplication of *pv1* or *tm-cc(at)*, the gold standard used to demonstrate that two genes share a common ancestor is to detect sequence similarity between them. To potentially facilitate the detection of homology between distantly related sequences, we used ML methods to reconstruct ancestral amniote PV1 and TM-CC(aT) protein sequences. It was not feasible to reconstruct reliable ancestral Tetherin sequences due to the large divergence. Thus, we searched for sequence similarity among extant and ancient PV1, TM-CC(aT) and Tetherin/TM-CC-GPI protein sequences using local alignment tools (BLAST).

While PV1 sequences were clearly homologous across all jawed vertebrates, TM-CC(aT) and Tetherin were more variable. In the case of Tetherin/TM-CC-GPI sequences, only the common protein architecture and position in genomes suggested that the proteins from mammals, birds, reptiles, fish and shark might share a common origin (Figures 6.7A and 6.9A). Conversely, TM-CC(aT) proteins from mammals, reptiles, avian, and coelacanth exhibited sequence similarity to each other, indicating an unambiguous common origin for vertebrate *tm-cc(at)* genes (Figure 6.9A). Attempts to detect homology between PV1 and TM-CC(aT) or Tetherin were statistically inconclusive, even when amniote ancestral sequences were used. Nevertheless, probabilistic models to detect distant homologs (HMMER3 (Finn et al., 2011)) found a significant hit (e-value =

2E-5) between coelacanth PV1 and a HMM profile of TM-CC(aT) proteins (Figures 6.9C and 6.10).

As noted above, the coelacanth has three *tm-cc(at)* genes and one *tetherin/tm-cc-gpi* gene in the *pv1–cilp2* interval (Figure 6.7A). Strikingly, the coelacanth Tetherin/TM-CC-GPI protein exhibited clear sequence similarity to the TM-CC(aT) proteins, particularly the TM-CC(aT)$_B$ protein encoded by the neighboring gene, unambiguously indicating a common ancestor (Figures 6.9B and C). Because coelacanth Tetherin/TM-CC-GPI is clearly a functional antiviral protein (Figures 6.6A and B) sequence similarity reveals that coelacanth *tetherin/tm-cc-gpi* shares a common ancestor with *tm-cc(at)* genes in diverse vertebrates.

Figure 6.9: Sequence similarity between PV1, TM-CC(aT) and Tetherin/TM-CC-GPI proteins.

(A) A heat map showing e-values of all combinations of reciprocal BLASTp analyses using the PV1, TM-CC(aT) and Tetherin/TM-CC-GPI proteins in this study. NS: non significant **(B)** Phylogenetic tree of divergent TM-CC(aT) protein sequences. ML tree was constructed using RAxML with 1000 bootstrap replicates and the alignment in (C). Numbers indicate bootstrap values. (**C)** Alignment of divergent TM-CC(aT) and Tetherin/TM-CC-GPI protein sequences from human (adapted from GenBank: XP_011526778.1), mouse (GenBank: XP_003945491.1), turkey (GenBank: XP_010723297.1), alligator (GenBank: KQL90195.1), turtle (adapted from GenBank: XP_008169839.1) and coelacanth (GenBank: XP_006001674.1 and XP_014347293.1, Gnomon prediction: 16424589.p and TM-CC(aT)$_A$ adapted from RNAseq reads on NW_005819727.1). Sequences spanning the TM, CC domains and GPI anchor are indicated. Residues that comprise the proline-rich domain in human and mouse TM-CC(aT) proteins are indicated in red.

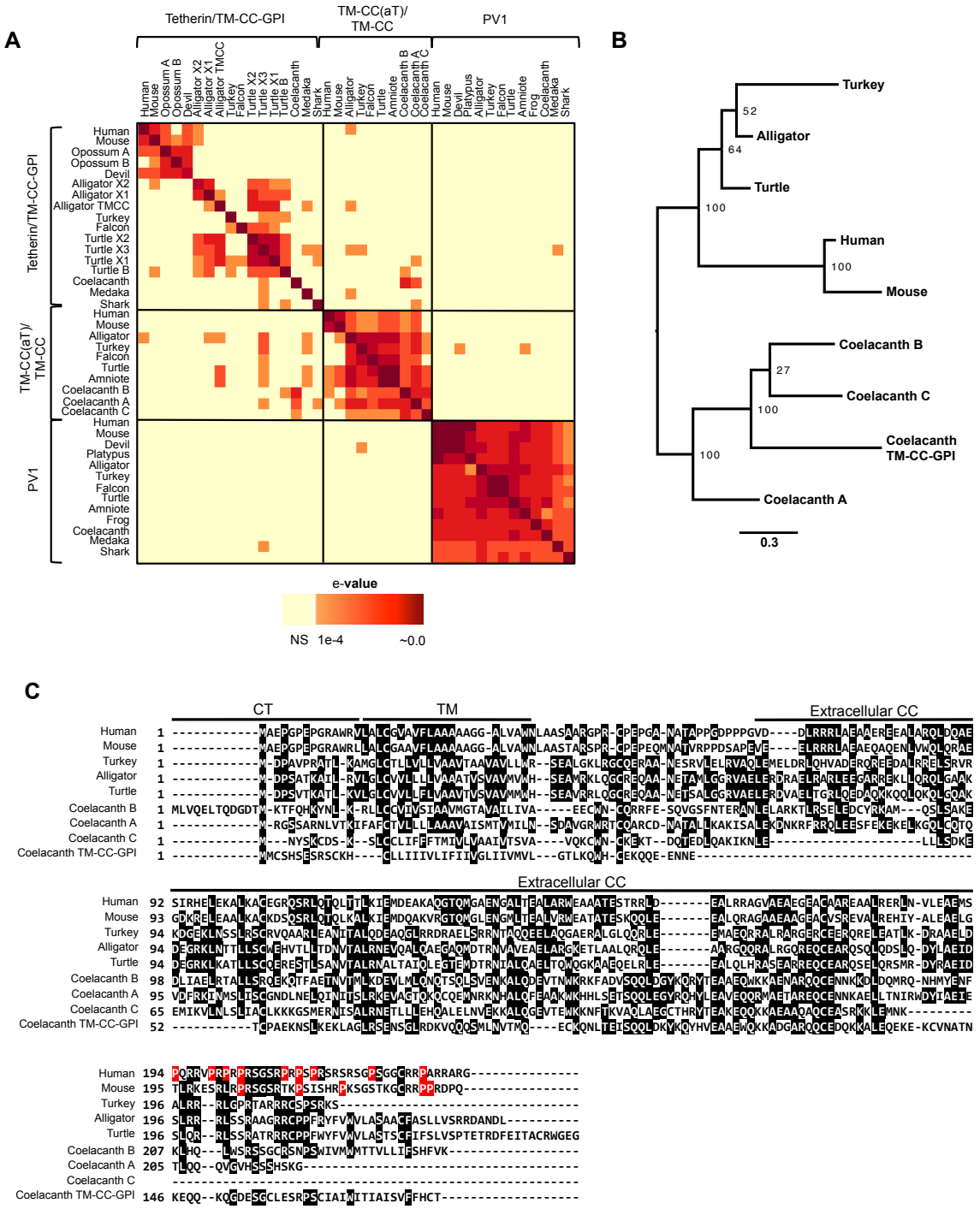


Figure 6.9: Sequence similarity between PV1, TM-CC(aT) and Tetherin/TM-CC-GPI proteins.

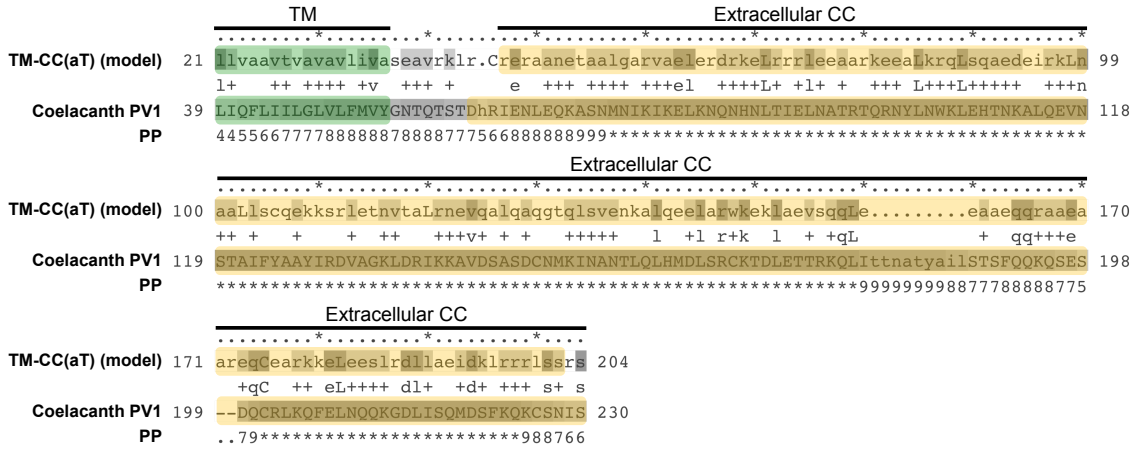


Figure 6.10: Sequence similarity between TM-CC(aT) and Coelacanth PV1 proteins.

Sequence alignment between coelacanth PV1 and a model inferred from a multiple sequence alignment (MSA) of TM-CC(aT) proteins (Figure 6.9C). Alignment built using HMMER3. Sequences spanning the transmembrane domain (TM) and coiled-coil domain (CC) are indicated. Conserved residues in the TM-CC(aT) MSA are shown in uppercase. A Dot (.) indicates the introduction of a gap in the model TM-CC(aT). Plus (+) signs indicate similar residues between the sequences. PP represents the posterior probability of each aligned residue (4 = 35-45%, 5 = 45-55%, … 9 = 85-95%, (*) = 95-100%).

Potential for tm-cc(at) to encode a GPI anchor in some species.

Although *tm-cc(at)* and *tetherin/tm-cc-gpi* genes share no sequence similarity in most vertebrates, and appear to have been placed under different types of selection in mammals, the finding that *tm-cc(at)* and *tetherin/tm-cc-gpi* (and possibly *pv1* (Figure 6.10)) are clear homologs in the coelacanth prompted us to ask how one might have arisen from the other. Because a key difference between these two proteins in mammals is the presence of a proline rich C-terminus in TM-CC(aT) versus a C-terminal GPI anchor in Tetherin, we inspected the annotated 3' exons of *tm-cc(at)* in non-mammalian species to determine how the GPI modification might has arisen. Notably, bona fide, near full length *tm-cc(at)* cDNA sequences from mouse and chicken differ in the number of exons, that contribute to sequence encoding the *tm-cc(at)* C-terminus. Specifically, mouse *tm-cc(at)* has four exons, with the fourth exon encoding a proline-rich C-terminal sequence (Figure 6.11A). Conversely the chicken *tm-cc(at)* fourth exon is truncated by splicing to a fifth exon encoding only two C-terminal amino acids (Figure 6.11A). Human *tm-cc(at)* is predicted to include a fifth exon that appends seven C-terminal amino acids that are absent in mouse *tm-cc(at)*. RNAseq data indicates that coelacanth *tm-cc(at)*$_A$ and *tm-cc(at)*$_B$ includes a fifth exon that encodes only two or one amino acid respectively, while the 3' end of the fourth exon in *tm-cc(at)*$_B$ encodes C-terminal sequences that have a high probability for a GPI modification, followed by a stop codon (Figure 6.11A). Thus, unlike

mammalian and avian TM-CC(aT) proteins, it is highly likely that coelacanth TM-CC(aT)$_B$ is GPI modified at its C-terminus, whether or not the fifth exon is used.

The *tm-cc(at)* gene in two reptile species (painted turtle and Chinese alligator) has sequences that potentially encode a hydrophobic amino acid-rich sequence immediately 3' to the fourth exon, while a potential fifth exon codes for 3-7 amino acids and a termination codon. We reasoned that, similar to the coelacanth variant, the hydrophobic amino acid-rich sequence in reptilian TM-CC(aT) proteins might confer GPI modification (Figure 6.11A). Due to a paucity of RNAseq data, it is unknown whether modern reptile *tm-cc(at)* genes encode four- or five-exon proteins or if they are GPI-modified at their C-termini. However, unlike their mammalian counterparts, reptile TM-CC(aT) genes have the potential to encode C-terminally GPI-modified proteins.

We constructed cDNAs expressing the coelacanth TM-CC(aT)$_B$ protein and the potential alternatively spliced versions of the alligator and turtle TM-CC(aT) proteins. The coelacanth TM-CC(aT)$_B$ proteins were not glycosylated and were poorly active, despite abundant expression (Figures 6.11B, C and D). However, the inclusion of the hydrophobic amino acid-rich sequence (i.e. GPI-modified) in the alligator and turtle TM-CC(aT) proteins conferred antiviral activity that was enhanced by the presence of the fifth exon (Figures 6.11B and C). Strikingly, both the four and five-exon versions of the turtle TM-CC(aT) proteins that contained the hydrophobic sequence potently inhibited the release of infectious HIV-1 particles by ~100-fold at the highest dose tested (Figures 6.11B and C).

These data suggest that *tm-cc(at)* genes in reptiles have the potential to encode GPI-modified proteins, with some isoforms exhibiting potent antiviral activity.

Summary

Tetherin encodes an antiviral protein with no known homologs. In this chapter, we describe scenarios by which this orphan gene may have arisen and evolved that exemplify how protein modularity, evolvability and robustness can create new functions and preserve them, despite sequence divergence due to genetic conflict. We find that Tetherin function is encoded by genes that exhibit no sequence similarity and share only a common architecture and location in modern genomes. Moreover, *tetherin* is part of a cluster of three potential sister genes that encode proteins of similar architecture, some variants of which exhibit antiviral activity while others can be endowed with antiviral activity following a simple modification. Only in a slowly evolving species (e.g. coelacanths) *tetherin* exhibits sequence similarity to one potential sister gene. We suggest that neofunctionalization, drift and positive selection lead to an intense sequence diversification among modern *tetherin* genes, and between *tetherin* and its sister genes, which obscured the ability to detect their homology.

Figure 6.11: Antiviral activity of non-mammalian TM-CC(aT) variants.

(A) Transcript structure and C-terminal protein sequences of potential alternatively spliced isoforms of *tm-cc(at)* in non-mammalian species. The TM, CC, proline-rich domains and hydrophobic patch are indicated in color. The omega site (underlined in blue) and specificity (1 – false positive rate) were predicted using PredGPI (Pierleoni et al., 2008). The number of RNAseq reads supporting the occurrence or absence of splicing events is indicated between exons. **(B)** Infectious virion yield measured using HeLa TZM-bl indicator cells following transfection of Vpu-deficient HIV-1 proviral plasmids along with plasmids expressing alternatively spliced isoforms of TM-CC(aT) proteins. RLU: relative light units. Data from 3 independent experiments. **(C)** Western blot analyses (anti-CA) of cell lysates and virions corresponding to (B). Numbers at the bottom represent virion CA protein levels relative to those obtained in the absence of an inhibitor. **(D)** Western blot analyses following transfection of Vpu-deficient HIV-1 proviral plasmids along with plasmids expressing HA-tagged four- or five-exon versions of TM-CC(aT) proteins. The cell lysates were probed with anti-HA antibodies.

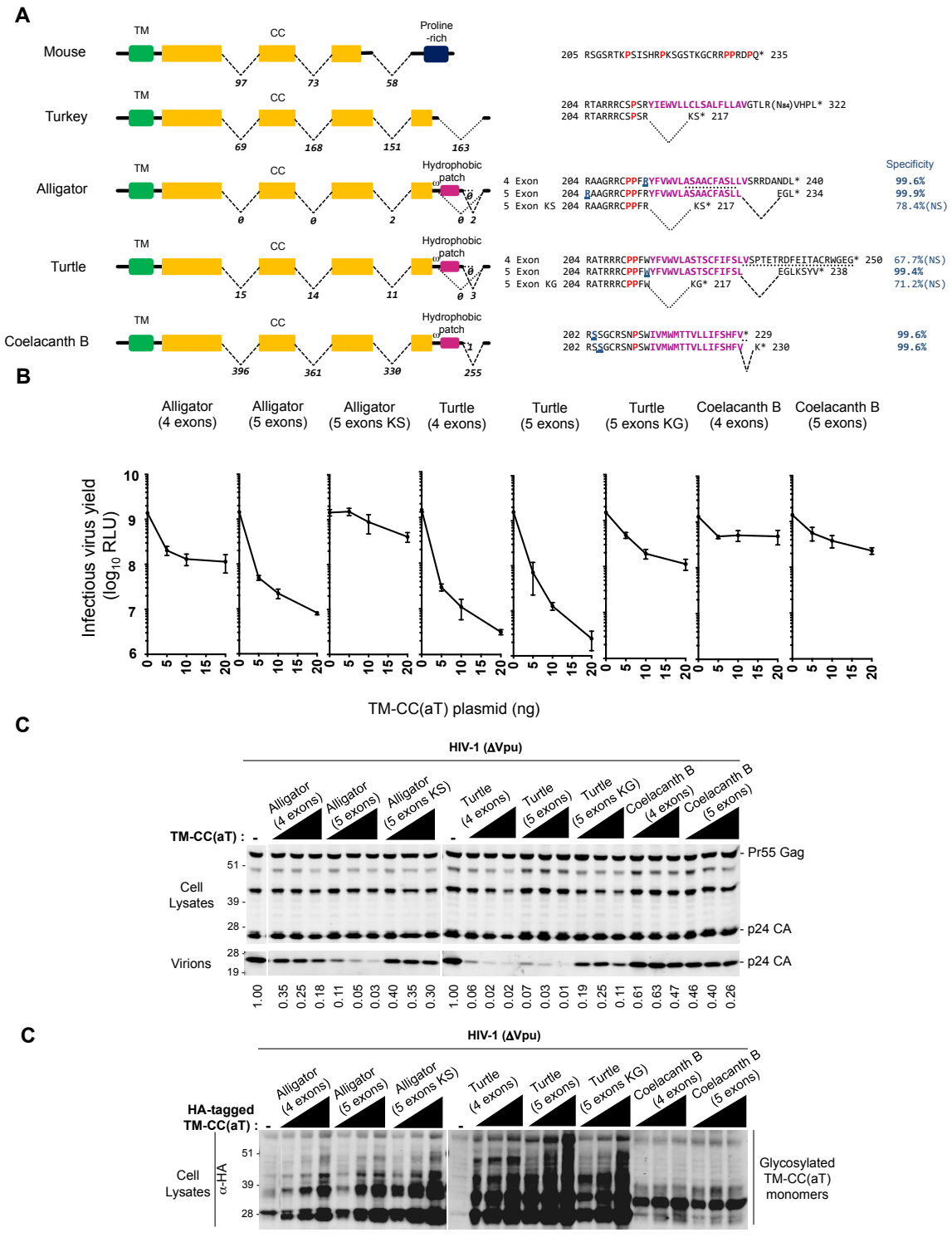


Figure 6.11: Antiviral activity of non-mammalian TM-CC(aT) variants.

Chapter VII. Discussion

In the preceding chapters we performed a comprehensive paleovrological study of selected ERV lineages and host defense mechanisms. We first developed a framework (DIGS) to heuristically explore the genomic landscape of ERVs through BLAST-based screens used in combination with strategically chosen reference datasets. We used DIGS to expand a previously produced phylogenetic screen based on RT in order to guide the recovery of proviral loci and the inference of ancestral genome sequences. This analysis showed the capability of DIGS to identify complex proviral structures despite the numerous mutational processes by which these sequences have evolved and revealed that most ERV lineages analyzed have a significant amount of mosaic species, which might represent true recombination events.

Although with no identifiable *env* gene, its high copy numbers in the mouse genome, its recent expansion and the opportunity to uncover a potential role in mouse development made MuERV-L an interesting candidate for ancestral reconstruction. We presented the successful resurrection of a ~2 MY old infectious ancestral MuERV-L sequence (ancML) by analysis of the “fossilized” MuERV-L elements in the mouse genome. These elements account for a particularly reduced set of all the integration events that occurred since its origin. Despite the probable random integration pattern of ancML (Figure 4.4D), the analysis of fixed MuERV-L elements showed that there is a significant selective pressure to eliminate integrations in genic regions and this pressure seems to

have being stronger for older elements (Figure 3.5). ERV-L elements differ from other highly abundant env-defective ERVs in that the *env* gene is completely absent and it is yet not clear if the *env* gene was lost in the founder element or if this ERV never had an *env* gene. Whichever the case, the bulk of the replication of these elements must have been entirely intracellular through retrotransposon-like mechanisms (Magiorkinis et al., 2012).

According to previous studies (Costas, 2003), and corroborated by ours, MuERV-L originated ~10 MYA after the *Rattus-Mus* split and underwent a very prolific expansion ~2 MYA in modern mouse species. In fact almost 65% of soloLTRs and MuERV-L proviruses identified in this study have an integration date below 3 MY. It is possible to speculate that the co-option of the MuERV-L LTR as a promoter for the genes involved in the zygotic genome activation might have also occurred ~2 MYA. Therefore, the sudden change in the transcriptional profile at an early developmental stage of the mouse embryo resulted in the highly productive expansion. Additionally, it is also possible that the embryonic environment could contribute to MuERV-L expression and replication, perhaps explaining the reduced set of cell lines in which ancML could replicate. Recent studies have shown that some ERV sequences have a fundamental role on mammalian development (Lavialle et al., 2013; Lu et al., 2014; Macfarlan et al., 2012; Wang et al., 2014). However, it yet remains to be determined if the presence of MuERV-L transcripts, proteins and virus-like particles at the two-cell stage play a role in early mouse development, or if this expression is merely a

byproduct of the LTR co-option. The generation of ancML certainly represents a unique opportunity to resolve this question and potentially uncover a new role for ERVs in mammalian development.

Currently there is no evidence that the ongoing expression of MuERV-L elements at the two-cell stage of the mouse embryo results in successful re-integration (Guallar et al., 2012). Consequently, either inactive MuERV-L elements recently lost the ability to be complemented *in-trans*, or an additional mechanism has recently evolved that inhibits re-integration of active MuERV-L elements. We observed that mouse IFN-α is able to inhibit ancML replication and therefore mount a successful antiviral response (Figure 4.5A). Expression analyses of the IFN-α receptor (IFN-α/β receptor composed of two subunits IFNAR1 and IFNAR2) during mouse development showed that *ifnar2* is poorly expressed at the two-cell stage whereas *ifnar1* (the other subunit) belongs to the top 30% most expressed genes at that developmental stage (Xie et al., 2010). It has previously been shown that IFNAR1 can mount a type I IFN signaling response in the absence of IFNAR2 (de Weerd et al., 2013). Therefore, it is possible that type I IFN responses might inhibit MuERV-L replication at the two-cell stage of the mouse embryo and have kept MuERV-L copy numbers under control.

We also observed that mouse APOBEC3, but not mouse MOV10 or SAMHD1, potently inhibits ancML replication (Figure 4.5B). However, mutational profiles of MuERV-L elements in the mouse genome failed to identify sufficient evidence for mA3-dependent hypermutation as a mechanism for inactivation (Figures 4.5C, D

and E), suggesting that the effects of hypermutation have not being critical for MuERV-L replication during its evolution. The expression profile of the mA3 during mouse development showed transcription at the two-cell stage of the mouse embryo with a ~3 fold increase at the four-cell stage becoming one of the top 30% most expressed genes at that developmental stage (Xie et al., 2010). The epigenetic silencing of MuERV-L after the two-cell stage makes improbable that the abundance of mA3 at the four-cell stage has had any effect on MuERV-L replication. Nevertheless, the presence of mA3 at the two-cell stage indicates that it is at least possible for mA3 to have acted on replicating MuERV-L elements through deaminase-independent mechanisms (MacMillan et al., 2013). Although there is no definitive answer as to what mechanism was responsible for the halt of MuERV-L replication, our results suggest that mA3, as well as a subset of IFN-stimulated genes, might have acted upon replicating MuERV-L elements and contributed to its extinction.

Despite its low copy number, another interesting candidate for ancestral reconstruction was HERV-T, due to the relative conservation of its *env* gene, the intriguing long leader sequence and the fact that no extant human gammaretrovirus is known. Analysis of HERV-T sequences in catarrhine primate genomes revealed that there are three classes of HERV-T elements, which possibly derived from independent infection and integration events, distinct expansion periods, or a mixture of both. Establishing with absolute certainty which of these scenarios was responsible for the existence of a particular class is

impossible. However, at least for HERV-T2 elements we hypothesize that they might have shared an early evolutionary history with HERV-T1, indicated by their clustering into a well supported monophyletic clade (Figure 3.1), as well as by their initial expansion times ~32 MYA (Figure 3.2). At some point along the way a subset of HERV-T2 elements were subjected to APOBEC3-mediated hypermutation, and subsequently expanded in apes, most likely through retrotransposon-like mechanisms. The oldest class, HERV-T3, originated previous to the old world monkey – ape split (~43 MY) and might have derived from HERV-T-like elements present in new world monkeys. These HERV-T-like elements share sequence similarity with the HERV-T3 coding sequence but not to the LTRs.

HERV-T elements show a striking long leader sequence that is unusual for gammaretroviruses and exogenous retroviruses in general. In fact the length of the HERV-T leader sequence is only exceeded by HERV-W with a leader sequence of 1937 nucleotides (Jurka et al., 2005). Although rare, several endogenous and exogenous retroviruses show an additional ORF in their leader sequence, such as the fish epsilonretrovirus WDSV (Holzschu et al., 1995) and HERV-H (Jern et al., 2005). Phylogenetic and statistical analyses of HERV-T3 and HERV-T1 ancestral sequences revealed an ORF previous to *gag* and spanning almost half of the leader sequence. The translated product of this *pre-gag* ORF encodes a putative transmembrane domain that shows no sequence similarity to any protein in public databases. Expression of a reconstructed

ancestral pre-Gag protein (HTpG) in virus producer cells shows a minor pro-viral effect specific to the MLV amphotropic envelope but not to the ecotropic. Attempts to determine if HTpG is a functional analogous of the gammaretroviral glycoGag have been futile. Although we cannot discard the possibility that the reconstructed sequence do not represent the real functional ancestor, the function of this ORF is yet unknown.

We additionally reconstructed a ~32 MY old infectious ancestral HERV-T3 envelope (ancHTenv) that is able to pseudotype MLV particles and infect a variety of primate, rodent and carnivore cell lines. This ancestral envelope pushes back the boundaries of time for paleovirological studies, representing the oldest retroviral protein "resurrected" so far. The ancHTenv sequence shows a predicted functional propeptide furin cleavage site, a signal peptide, a transmembrane domain and a fusion peptide. Its fusogenic ability was further corroborated by the ability to form syncytia in 293T cells. This reconstruction was certainly facilitated by the relative conservation of the *env* sequences of HERV-T elements, which suggest that this ERV expanded to moderate copy numbers by reinfection of germ line cells using a functional envelope. By expressing a 293T-derived cDNA library in a non-susceptible cell line (chicken DF-1 cells) we were able to identify MOT1 as the receptor used by ancHTenv. This receptor has 12 transmembrane domains and facilitates solute (monocabroxylates) transport across the plasma membrane, similarly to other gammaretroviruses receptors (Figure 1.4A). This observation highlights the close phylogenetic relation between

HERV-T and gammaretroviruses, and supports the possibility of HERV-T being itself an extinct human gammaretrovirus.

MOT1 is part of the solute carrier 16 (SLC16) family of proteins that contains 14 members, four of which (MOT1-MOT4) have been identified as true monocarboxylate transporters (Halestrap and Wilson, 2012). MOT1 is widely expressed in human tissues, including thyroid, ovary and testes (Consortium et al., 2014), and its sequence is highly conserved among animal species, which might explain the broad tissue and species tropism observed for ancHTenv. Moreover, there are a couple of residues in the extracellular loops of MOT1 that are conserved between species whose cells were permissible to infection of ancHTenv containing particles (primates, Chinese hamster and dog) but not among species whose cells were non-permissible (mouse, rat and chicken) (Figures 5.6A and 7.1). These residues might be responsible for the reduced ability of ancHTenv pseudotyped particles to infect DF-1, NIH3T3 and possibly Rat2 cells. Further experiments are needed to corroborate the ancHTenv–MOT1 interaction and to address its possible determinants.

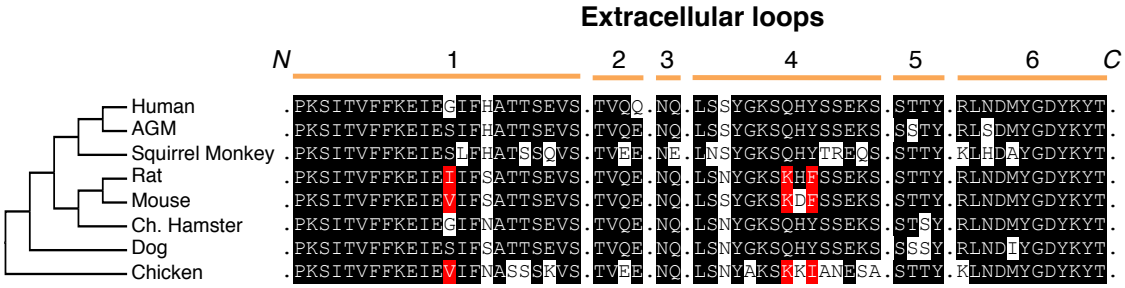


Figure 7.1: Sequence of the extracellular domains of MOT1 in species tested for susceptibility to infection by ancHTenv.

Multiple sequence alignment for the extracellular domains of MOT1 in selected species. Residues in black are conserved by at least 30% of the sequences. Residues colored in red are those ones that differ exclusively in mouse, rat and chicken and might confer resistance to ancHTenv. The number of the extracellular loops is indicated on top. Dots in the alignment represent continuity of the MOT1 sequence. N and C termini are indicated in italics. AGM: African Green Monkey. Ch. Hamster: Chinese hamster.

Remarkably, we identified a HERV-T proviral locus with a complete protein-coding *env* gene in the human genome (HsaHTenv), which is unable to pseudotype MLV particles, but shows a potent and specific inhibitory effect against the functional ancestral HERV-T envelope (ancHTenv). Although it is documented that HsaHTenv is well expressed in healthy thyroid tissue (but also expressed in others (de Parseval et al., 2003)), we were unable to amplify it by RT-PCR from cancerous cells lines of thyroid origin. The thyroid gland is a complex tissue composed of mainly two types of cells, follicular epithelial cells and parafollicular cells, surrounded by highly vascular connective tissue (Fawcett and Jensh, 1997). Therefore, the expression of HsaHTenv might be specific to one healthy type of cell and could have been misrepresented in the cancerous cell lines tested. One mode of action by which this antiviral *env* gene might act is by receptor interference, where the product of this gene (HsaHTenv) could interact with the receptor used by ancHTenv (MOT1), thereby preventing its interaction with the functional ancestral envelope and the subsequent infection (Figure 7.2). During its transport to the plasma membrane, MOT1 could be sequestered into internal vesicles by interacting with HsaHTenv, reducing its surface expression and inhibiting infection by ancHTenv pseudotype particles (Figure 7.2B). Alternatively HsaHTenv might be expressed on the cell surface and interact with MOT1, resulting in the blockage of available receptors and the inhibition of ancHTenv infection (Figure 7.2C). It is unlikely that HsaHTenv (or some portion) has to be secreted in order to interact with MOT1, due to the

presence of a transmembrane domain and the lack of a functional furin cleavage site.

Infection by ancHTenv pseudotyped MLV particles is almost 3 fold lower in cells expressing the most recent ancestor of HsaHTenv (ancHsaHTenv) compared to the human copy. Suggesting, that the inhibitory function of this *env* gene was evolved shortly after its integration in the ancestral hominid genome ~13-19 MYA. This *env* ORF is complete in humans and orangutans but an extra cytidine in a 7-citidine low complexity region in the gorilla ortholog results in a premature termination. However, the otherwise complete *env* sequence continues undisrupted in a different reading frame (Figure 5.10B). It is possible that this frameshift insertion occurred recently and the remains of the ORF have not had time to diverge. Alternatively, the genomic sequence of the gorilla might not be completely accurate at this low complexity region and a complete *env* ORF is still present in gorillas. Analysis of all hominid *env* sequences shows that they are evolving slower than the rest of the provirus and they have been remarkably refractory to non-sense mutations, in contrast to their cognate *gag-pol* sequences. Although, *in silico* evolution simulations showed a small but significant amount of sequences retaining coding potential (2-8% of simulated sequences), these results are not taking into account the further inactivating mutations derived from insertion/deletion events fixed by genetic drift. With all of these observations taken into consideration it is unlikely that this *env* gene has retained its coding potential in the absence of any selective pressures (neutral

evolution). Indicating its probable co-option as a hominid restriction factor for HERV-T and/or related viruses.

Examples of ERV *env* genes mediating receptor interference have been documented in chickens (Robinson et al., 1981), sheep (Spencer et al., 2003), mice (Gardner et al., 1991; Odaka et al., 1980; Wu et al., 2005) and cats (Ito et al., 2013). Therefore, it seems that this strategy represents an efficient way to inhibit viral infections to an extend that it has been recurrently adapted at different times by different hosts to deal with viral infections. Further experiments are needed to address if this is the mode of action of HsaHTenv, which would represent the first documented case for a human restriction factor to be derived from an ERV *env* gene.

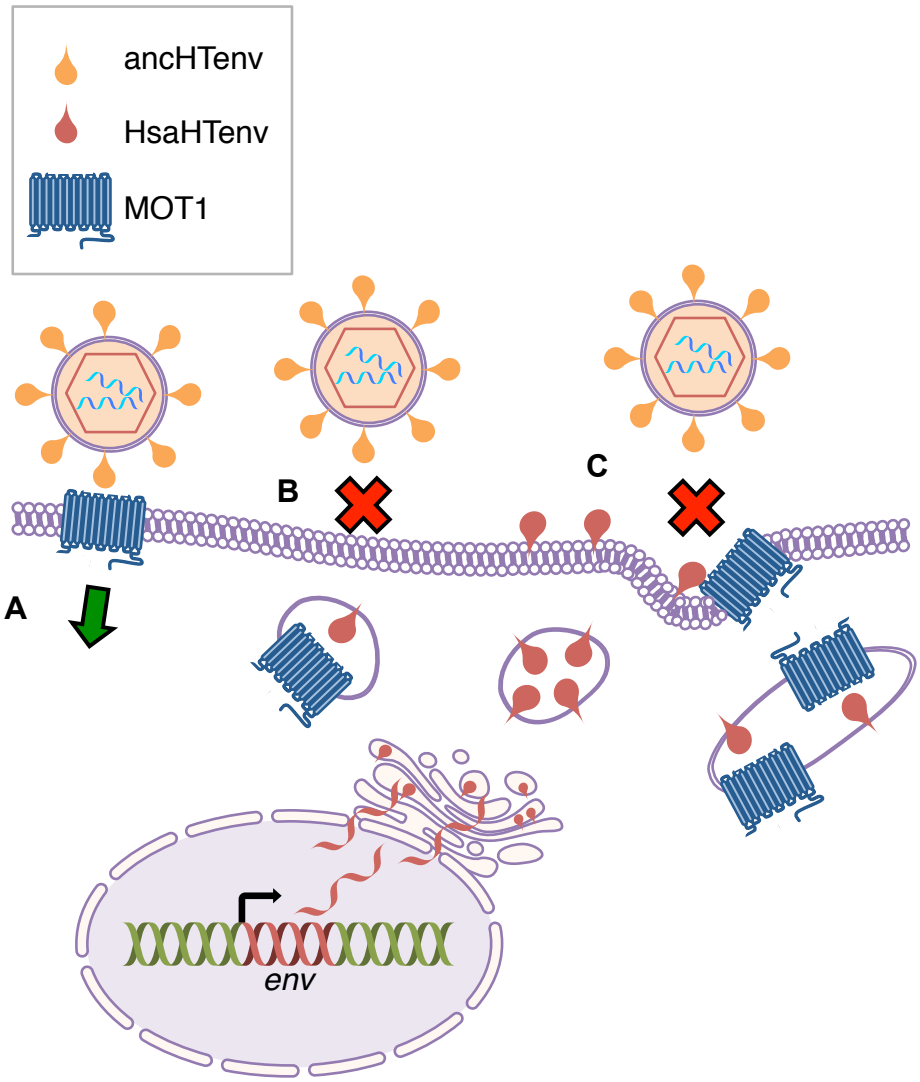


Figure 7.2: Model for the possible antiviral effect of HsaHTenv through receptor interference in humans.

(A) ancHTenv uses MOT1 as a receptor to mediate viral entry. **(B)** MOT1 could be sequestered into internal vesicles by interaction with the product of the protein-coding *env* gene in the human genome (HsaHTenv), thereby preventing its surface expression and inhibiting infection by ancHTenv pseudotype particles. **(C)** Alternatively HsaHTenv might be expressed in the cell surface and interact with MOT1, resulting in the blockage of available receptors and inhibiting infection by ancHTenv pseudotype particles. Afterwards MOT1 could be internalized in to intracellular vesicles and target for degradation.

Based on the integration dates estimated for HERV-T3 soloLTRs (Figure 3.2), we infer that the end of the major expansion period for this HERV-T class (10-15 MYA) overlaps with the estimated dates for the acquisition of the antiviral *env* copy (13-19 MYA). These dates also coincide with the estimated ones based on paired LTR comparisons (Figure 5.1), except for a pair of 8 MY old proviruses specific to macaques, which do not encode for this antiviral *env* gene. This circumstantial, but suggestive observation, argues in favor for the co-option of the protein-coding *env* gene to protect against HERV-T3 infection, resulting in the halt of its expansion and perhaps even its extinction. Under this hypothesis it would be unclear how HERV-T1 viruses escaped the action of this *env* gene and continued to expand in hominids (estimated by the dates of its soloLTRs) (Figure 3.2). However, *env* containing HERV-T1 loci have dates that would also overlap with the acquisition of the antiviral *env* copy (Figure 5.1). Thus, it might still be possible that the few remaining HERV-T1 elements continued expanding through mechanisms independent of infection. As seductive as this hypothesis might sound the idea is just merely speculative.

As paleovirology also studies the evolution of the host defense mechanisms that have been shaped by past retroviral infections, we investigated the origins and evolution of Tetherin, an orphan antiviral protein with no known homologs. The lack of sequence similarity to other proteins and the inaccuracies of the assembly and annotation of some vertebrate genomes complicated this endeavor. Nevertheless, our findings strongly suggest that the genesis of *tetherin* occurred

by duplication and neofunctionalization of an ancestor of a neighboring gene encoding a TM-CC protein (*pv1* and/or *tm-cc(at))*. This conclusion is based on the findings that: (i) genes encoding TM-CC proteins constitute ~1% of all genes and most of those that are arranged contiguously obviously share a common ancestor; (ii) *pv1*, *tm-cc(at)* and *tetherin/tm-cc-gpi* are proximal in many modern species and share a similar exon-intron structure; (iii) PV1 and especially TM-CC(aT) are able to trap virions following a simple manipulation that could have plausibly been acquired to enable GPI modification; (iv) in the coelacanth, TM-CC(aT) and Tetherin/TM-CC-GPI are obvious homologs; and (v) some modern *tm-cc(at)* genes have the potential to encode GPI modified proteins and exhibit antiviral activity.

It is clear that the detection of distant homologs is still a challenge for current sequence similarity methods. Nevertheless, probabilistic models to detect distant homologs (HMMER3 (Finn et al., 2011)) found a significant hit (e-value = 2E-5) between coelacanth PV1 and a HMM profile of TM-CC(aT) proteins (Figures 6.9C and 6.10). This finding, coupled with our observation that the coelacanth Tetherin/TM-CC-GPI protein shares significant sequence similarity with the TM-CC(aT)$_B$ protein, might reflect the slow neutral substitution rate in the coelacanth (Amemiya et al., 2013). However, we were unable to detect sequence similarity between Tetherin, or TM-CC(aT), and PV1 in another apparently slowly evolving species such as turtle (Shaffer et al., 2013), although there are residual levels of similarity between turtle Tetherin and TM-CC(aT) (Figure 6.9A).

Presumably, following acquisition of sequences specifying a GPI anchor, nascent Tetherin/TM-CC-GPI sequences were optimized in various species through positive selection pressures imposed by past viruses in order to enhance its potency and/or to avoid viral antagonists. Therefore, while it is intuitive to expect sequence similarity between proteins that have a common ancestor, our findings reveal that the lack of sequence similarity between *tetherin* genes in most species should be expected since initial genetic drift and subsequent positive selection for some fraction of ~450 MYA, appeared to have erased sequence similarity within its orthologs in other species to nearly the same degree as its sister genes, ultimately resulting in extreme diversity of Tetherin/TM-CC-GPI proteins and its previous classification as an orphan gene in modern vertebrates.

In the absence of sequence similarity, the occurrence of TM-CC-GPI protein-encoding genes in syntenic vertebrate genomic positions neighboring other TM-CC protein-encoding genes is the strongest single piece of evidence that modern *tetherin/tm-cc-gpi* genes share a common ancestor. The conservation and essential function of *pv1* strongly suggests that it was present in the vertebrate ancestor, with one or two duplications of it giving rise to *tm-cc(at)* and/or *tetherin/tm-cc-gpi* in the various vertebrate lineages. The strong antiviral effect of some GPI-modified TM-CC(aT) proteins, the homology between coelacanth TM-CC(aT)$_B$ and functional Tetherin/TM-CC-GPI proteins, and the shared gene structure suggest that most modern Tetherin proteins arose from TM-CC(aT)-like proteins. One possible model for the genesis of *tetherin/tm-cc-gpi* genes is that

pv1 duplicated first into *pv1–tm-cc(at).* Subsequent *tm-cc(at)* duplication yielded *pv1–tm-cc(at)–tetherin/tm-cc-gpi* resulting in a single common *tetherin/tm-cc-gpi* ancestor that was derived from *tm-cc(at)* prior to the division of sharks from other jawed vertebrate lineages (Figure 7.3A). In this model, gene loss and rearrangement events in sharks, ray-finned fish and amphibians would give the modern genome configurations, with sequence homology between *tetherin/tm-cc-gpi* and the ancestral *tm-cc(at)* detected only in the coelacanth lineage. Our finding that all vertebrate Tetherin/TM-CC-GPI proteins tested exhibit potent antiviral activity fits in this model and suggest that *tetherin/tm-cc-gpi* neofunctionalization occurred early during vertebrate evolution, prior to the division of sharks from other vertebrate lineages.

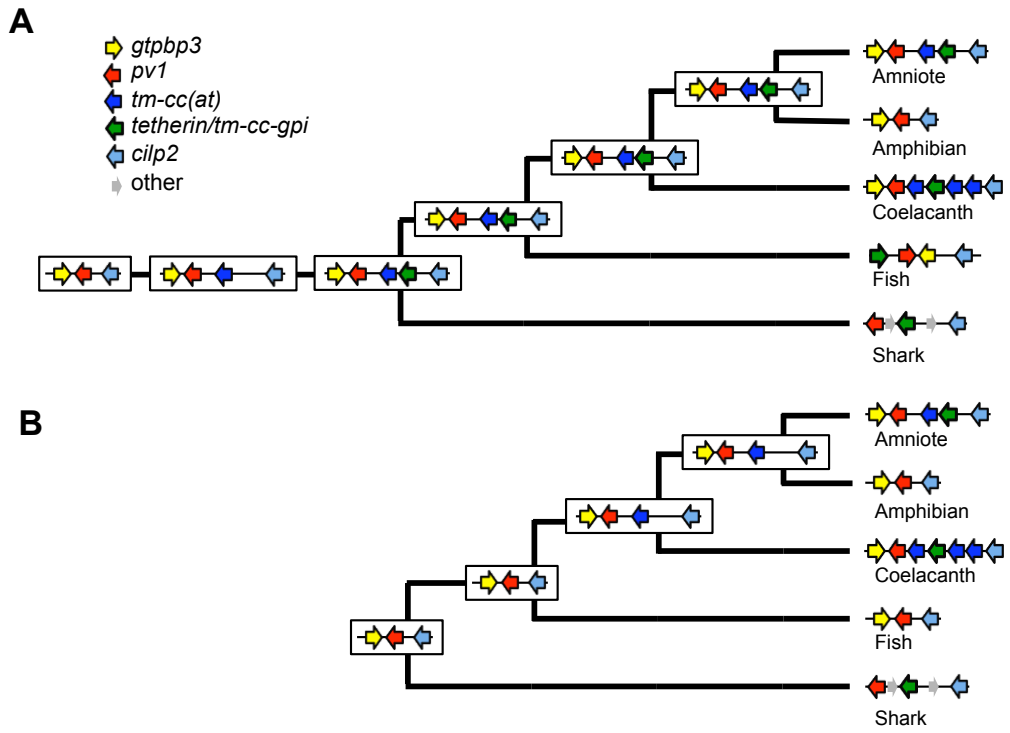


Figure 7.3: Possible evolutionary scenarios for the emergence of *tetherin/tm-cc-gpi* genes in the *pv1–cilp2* locus.

(A) *tetherin/tm-cc-gpi* originated once, prior to the division of sharks from other jawed vertebrate lineages via sequential duplications of *pv1* and *tm-cc(at)*. **(B)** *tetherin/tm-cc-gpi* originated independently in multiple vertebrate lineages via duplications of *pv1* and *tm-cc(at)*. This figure was kindly provided by Dr. Paul D. Bieniasz.

However, it is not necessarily the case that a single ancestral *tetherin/tm-cc-gpi* gene arose in vertebrates on one occasion and gave rise to all modern Tetherin/TM-CC-GPI proteins. Another possible model would be one where *pv1* duplicated into *pv1–tm-cc(at)* in the sarcopterygiian ancestor (lobe-finned fish and terrestrial vertebrates), after the separation of sharks and ray-finned fish. Thereafter, an independent *pv1* duplication in sharks, and separate *tm-cc(at)* duplication events in coelacanths and various amniote lineages yielded *tetherin/tm-cc-gpi* genes in their modern configurations (Figure 7.3B). In this model, it is implausible that *tm-cc-gpi* genes could have arisen on multiple occasions in nearly the same genomic location, unless they are indeed duplications of proximal TM-CC genes (*pv1* or *tm-cc(at)*).

While PV1 is essential for mouse viability, and its function in the generation of diaphragms in caveolae, fenestrae and transendothelial channels has been elucidated (Stan et al., 2012), the function of TM-CC(aT) proteins is unknown. RNA-Seq data suggests that it is poorly or rarely expressed in humans, however we were able to detect a spliced mRNA species in 7-day old mouse embryos. The apparent absence of *tm-cc(at)* in opossums, as well as in fish and shark lineages suggests that it does not play an essential role in the life of cells. Notably, some versions of *tm-cc(at)*, like *pv1*, encode a proline rich-C terminus, others are predicted to encode a GPI anchor, and still others have no defining C-terminal feature. It is then plausible that *tm-cc(at)* has different functions in different species. It is also possible that *tm-cc(at)* was, or might still be a

functional *tetherin* in some species, particularly since some reptile *tm-cc(at)* genes have the potential to encode a protein with potent antiviral activity. Our results suggest a key role for the loss or acquisition of splicing signals in the genesis of Tetherin proteins encoding the critical C-terminal GPI anchor.

To our knowledge, the *pv1–tm-cc(at)–tetherin* cluster is the only known gene triplet whose members share structural, but not sequence similarity with their sister genes, and have unrelated functions. However, Tetherin is unusual in that its biological function can be attributed largely to its unique TM-CC-GPI dual-anchored topology, rather than to its specific amino acid sequence. Therefore, the modular structure and rather simple function of Tetherin appears to have resulted in an unusual combination of extreme robustness and evolvability that first led to its genesis from a protein of unrelated function, and then enabled it to adapt to evade viral antagonists while maintaining antiviral activity. These findings highlight the key role of genomic duplication as a raw material for genetic innovation. Moreover, predicted structure comparisons, synteny and functional analyses might serve as valuable approaches for revealing the evolutionary history of other orphan genes.

In the present work we have identified the ERV diversity in primate and murine genomes, reconstruct ancestral infectious retroviruses and characterized how past viral infections have driven the evolution of host antiviral proteins. Overall, this study highlights the stunning potential of the retroviral “fossil record” to

uncover new biological processes and how the combination of *in silico* and *in vitro* analyses can answer profound evolutionary questions.

Bibliography

Aaronson, S.A., Hartley, J.W., and Todaro, G.J. (1969). Mouse leukemia virus: "spontaneous" release by mouse embryo cells after long-term in vitro cultivation. Proc Natl Acad Sci U S A *64*, 87-94.

Abudu, A., Wang, X., Dang, Y., Zhou, T., Xiang, S.H., and Zheng, Y.H. (2012). Identification of molecular determinants from Moloney leukemia virus 10 homolog (MOV10) protein for virion packaging and anti-HIV-1 activity. The Journal of biological chemistry *287*, 1220-1228.

Aiewsakun, P., and Katzourakis, A. (2015). Endogenous viruses: Connecting recent and ancient viral evolution. Virology *479-480*, 26-37.

Albritton, L.M., Kim, J.W., Tseng, L., and Cunningham, J.M. (1993). Envelope-binding domain in the cationic amino acid transporter determines the host range of ecotropic murine retroviruses. J Virol *67*, 2091-2096.

Albritton, L.M., Tseng, L., Scadden, D., and Cunningham, J.M. (1989). A putative murine ecotropic retrovirus receptor gene encodes a multiple membrane-spanning protein and confers susceptibility to virus infection. Cell *57*, 659-666.

Alkhatib, G., Combadiere, C., Broder, C.C., Feng, Y., Kennedy, P.E., Murphy, P.M., and Berger, E.A. (1996). CC CKR5: a RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1. Science *272*, 1955-1958.

Amemiya, C.T., Alfoldi, J., Lee, A.P., Fan, S., Philippe, H., Maccallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N.*, et al.* (2013). The African coelacanth genome provides insights into tetrapod evolution. Nature *496*, 311-316.

Anderson, M.M., Lauring, A.S., Burns, C.C., and Overbaugh, J. (2000). Identification of a cellular cofactor required for infection by feline leukemia virus. Science *287*, 1828-1830.

Andrew, A.J., Miyagi, E., Kao, S., and Strebel, K. (2009). The formation of cysteine-linked dimers of BST-2/tetherin is important for inhibition of HIV-1 virus release but not for sensitivity to Vpu. Retrovirology *6*, 80.

Arhel, N.J., Souquere-Besse, S., Munier, S., Souque, P., Guadagnini, S., Rutherford, S., Prevost, M.C., Allen, T.D., and Charneau, P. (2007). HIV-1 DNA Flap formation promotes uncoating of the pre-integration complex at the nuclear pore. EMBO J *26*, 3025-3037.

Arjan-Odedra, S., Swanson, C.M., Sherer, N.M., Wolinsky, S.M., and Malim, M.H. (2012). Endogenous MOV10 inhibits the retrotransposition of endogenous retroelements but not the replication of exogenous retroviruses. Retrovirology *9*, 53.

Arnaud, F., Black, S.G., Murphy, L., Griffiths, D.J., Neil, S.J., Spencer, T.E., and Palmarini, M. (2010). Interplay between ovine bone marrow stromal cell antigen 2/tetherin and endogenous retroviruses. J Virol *84*, 4415-4425.

Arnaud, F., Caporale, M., Varela, M., Biek, R., Chessa, B., Alberti, A., Golder, M., Mura, M., Zhang, Y.P., Yu, L.*, et al.* (2007). A paradigm for virus-host coevolution: sequential counter-adaptations between endogenous and exogenous retroviruses. PLoS Pathog *3*, e170.

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

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

Barr, S.D., Leipzig, J., Shinn, P., Ecker, J.R., and Bushman, F.D. (2005). Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. J Virol *79*, 12035-12044.

Battini, J.L., Danos, O., and Heard, J.M. (1998). Definition of a 14-amino-acid peptide essential for the interaction between the murine leukemia virus amphotropic envelope glycoprotein and its receptor. J Virol *72*, 428-435.

Battini, J.L., Heard, J.M., and Danos, O. (1992). Receptor choice determinants in the envelope glycoproteins of amphotropic, xenotropic, and polytropic murine leukemia viruses. J Virol *66*, 1468-1475.

Battini, J.L., Rasko, J.E., and Miller, A.D. (1999). A human cell-surface receptor for xenotropic and polytropic murine leukemia viruses: possible role in G protein-coupled signal transduction. Proc Natl Acad Sci U S A *96*, 1385-1390.

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

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

Benit, L., De Parseval, N., Casella, J.F., Callebaut, I., Cordonnier, A., and Heidmann, T. (1997). Cloning of a new murine endogenous retrovirus, MuERV-L,

with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. J Virol *71*, 5652-5657.

Benit, L., Dessen, P., and Heidmann, T. (2001). Identification, phylogeny, and evolution of retroviral elements based on their envelope genes. J Virol *75*, 11709-11719.

Benit, L., Lallemand, J.B., Casella, J.F., Philippe, H., and Heidmann, T. (1999). ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. J Virol *73*, 3301-3308.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2009). GenBank. Nucleic Acids Res *37*, D26-31.

Berger, E.A., Murphy, P.M., and Farber, J.M. (1999). Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. Annu Rev Immunol *17*, 657-700.

Bergthorsson, U., Andersson, D.I., and Roth, J.R. (2007). Ohno's dilemma: evolution of new genes under continuous selection. Proc Natl Acad Sci U S A *104*, 17004-17009.

Berlioz, C., and Darlix, J.L. (1995). An internal ribosomal entry mechanism promotes translation of murine leukemia virus gag polyprotein precursors. J Virol *69*, 2214-2222.

Best, S., Le Tissier, P., Towers, G., and Stoye, J.P. (1996). Positional cloning of the mouse retrovirus restriction gene Fv1. Nature *382*, 826-829.

Bieniasz, P.D. (2006). Late budding domains and host proteins in enveloped virus release. Virology *344*, 55-63.

Bieniasz, P.D. (2009). The cell biology of HIV-1 virion genesis. Cell Host Microbe *5*, 550-558.

Bishop, K.N., Holmes, R.K., Sheehy, A.M., Davidson, N.O., Cho, S.J., and Malim, M.H. (2004). Cytidine deamination of retroviral DNA by diverse APOBEC proteins. Curr Biol *14*, 1392-1396.

Blaise, S., de Parseval, N., Benit, L., and Heidmann, T. (2003). Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci U S A *100*, 13013-13018.

Blanco-Melo, D., Venkatesh, S., and Bieniasz, P.D. (2012). Intrinsic cellular defenses against human immunodeficiency viruses. Immunity *37*, 399-411.

Blond, J.L., Lavillette, D., Cheynet, V., Bouton, O., Oriol, G., Chapel-Fernandes, S., Mandrand, B., Mallet, F., and Cosset, F.L. (2000). 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 *74*, 3321-3329.

Blusch, J.H., Haltmeier, M., Frech, K., Sander, I., Leib-Mosch, C., Brack-Werner, R., and Werner, T. (1997). Identification of endogenous retroviral sequences based on modular organization: proviral structure at the SSAV1 locus. Genomics *43*, 52-61.

Bohnlein, S., Hauber, J., and Cullen, B.R. (1989). Identification of a U5-specific sequence required for efficient polyadenylation within the human immunodeficiency virus long terminal repeat. J Virol *63*, 421-424.

Breitbart, M., and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? Trends Microbiol *13*, 278-284.

Briggs, J.A., Wilk, T., and Fuller, S.D. (2003). Do lipid rafts mediate virus assembly and pseudotyping? J Gen Virol *84*, 757-768.

Brudno, M., Malde, S., Poliakov, A., Do, C.B., Couronne, O., Dubchak, I., and Batzoglou, S. (2003). Glocal alignment: finding rearrangements during alignment. Bioinformatics *19 Suppl 1*, i54-62.

Buiser, R.G., Bambara, R.A., and Fay, P.J. (1993). Pausing by retroviral DNA polymerases promotes strand transfer from internal regions of RNA donor templates to homopolymeric acceptor templates. Biochim Biophys Acta *1216*, 20-30.

Burdick, R., Smith, J.L., Chaipan, C., Friew, Y., Chen, J., Venkatachari, N.J., Delviks-Frankenberry, K.A., Hu, W.S., and Pathak, V.K. (2010). P body-associated protein Mov10 inhibits HIV-1 replication at multiple stages. J Virol *84*, 10241-10253.

Busnadiego, I., Kane, M., Rihn, S.J., Preugschas, H.F., Hughes, J., Blanco-Melo, D., Strouvelle, V.P., Zang, T.M., Willett, B.J., Boutell, C*., et al.* (2014). Host and viral determinants of Mx2 antiretroviral activity. J Virol *88*, 7738-7752.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Campbell, E.M., and Hope, T.J. (2015). HIV-1 capsid: the multifaceted key player in HIV-1 infection. Nat Rev Microbiol *13*, 471-483.

Campbell, S., and Vogt, V.M. (1997). In vitro assembly of virus-like particles with Rous sarcoma virus Gag deletion mutants: identification of the p10 domain as a morphological determinant in the formation of spherical particles. J Virol *71*, 4425-4435.

Carbone, L., Harris, R.A., Gnerre, S., Veeramah, K.R., Lorente-Galdos, B., Huddleston, J., Meyer, T.J., Herrero, J., Roos, C., Aken, B.*, et al.* (2014). Gibbon genome and the fast karyotype evolution of small apes. Nature *513*, 195-201.

Champoux, J.J., and Schultz, S.J. (2009). Ribonuclease H: properties, substrate specificity and roles in retroviral reverse transcription. FEBS J *276*, 1506-1516.

Chatterji, U., Bobardt, M.D., Stanfield, R., Ptak, R.G., Pallansch, L.A., Ward, P.A., Jones, M.J., Stoddart, C.A., Scalfaro, P., Dumont, J.M.*, et al.* (2005). Naturally occurring capsid substitutions render HIV-1 cyclophilin A independent in human cells and TRIM-cyclophilin-resistant in Owl monkey cells. The Journal of biological chemistry *280*, 40293-40300.

Choi, J., Ryoo, J., Oh, C., Hwang, S., and Ahn, K. (2015). SAMHD1 specifically restricts retroviruses through its RNase activity. Retrovirology *12*, 46.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083-1087.

Coffin, J.M., Hughes, S.H., and Varmus, H. (1997). Retroviruses (Plainview, N.Y.: Cold Spring Harbor Laboratory Press).

Connor, R.I., Sheridan, K.E., Ceradini, D., Choe, S., and Landau, N.R. (1997). Change in coreceptor use correlates with disease progression in HIV-1--infected individuals. J Exp Med *185*, 621-628.

Consortium, C.S.A. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. Nature *437*, 69-87.

Consortium, S.T.D., Williams, A.L., Jacobs, S.B., Moreno-Macias, H., Huerta-Chagoya, A., Churchhouse, C., Marquez-Luna, C., Garcia-Ortiz, H., Gomez-Vazquez, M.J., Burtt, N.P.*, et al.* (2014). Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature *506*, 97-101.

Corbin, A., and Darlix, J.L. (1996). Functions of the 5' leader of murine leukemia virus genomic RNA in virion structure, viral replication and pathogenesis, and MLV-derived vectors. Biochimie *78*, 632-638.

Cordonnier, A., Casella, J.F., and Heidmann, T. (1995). Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. J Virol *69*, 5890-5897.

Costas, J. (2003). Molecular characterization of the recent intragenomic spread of the murine endogenous retrovirus MuERV-L. J Mol Evol *56*, 181-186.

Craigie, R., and Bushman, F.D. (2014). Host Factors in Retroviral Integration and the Selection of Integration Target Sites. Microbiol Spectr *2*.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

Cui, J., and Holmes, E.C. (2012). Endogenous lentiviruses in the ferret genome. J Virol *86*, 3383-3385.

Dalgleish, A.G., Beverley, P.C., Clapham, P.R., Crawford, D.H., Greaves, M.F., and Weiss, R.A. (1984). The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. Nature *312*, 763-767.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. Nat Methods *9*, 772.

Daugherty, M.D., and Malik, H.S. (2012). Rules of engagement: molecular insights from host-virus arms races. Annu Rev Genet *46*, 677-700.

de Parseval, N., Lazar, V., Casella, J.F., Benit, L., and Heidmann, T. (2003). Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins. J Virol *77*, 10414-10422.

de Weerd, N.A., Vivian, J.P., Nguyen, T.K., Mangan, N.E., Gould, J.A., Braniff, S.J., Zaker-Tabrizi, L., Fung, K.Y., Forster, S.C., Beddoe, T.*, et al.* (2013). Structural basis of a unique interferon-beta signaling axis mediated via the receptor IFNAR1. Nat Immunol *14*, 901-907.

Debyser, Z., Christ, F., De Rijck, J., and Gijsbers, R. (2015). Host factors for retroviral integration site selection. Trends Biochem Sci *40*, 108-116.

DeFranco, A.L., Locksley, R.M., and Robertson, M. (2007). Immunity : the immune response in infectious and inflammatory disease (London

Sunderland, MA: New Science Press ;

Sinauer Associates).

Delchambre, M., Gheysen, D., Thines, D., Thiriart, C., Jacobs, E., Verdin, E., Horth, M., Burny, A., and Bex, F. (1989). The GAG precursor of simian immunodeficiency virus assembles into virus-like particles. EMBO J *8*, 2653-2660.

Delwart, E.L., and Panganiban, A.T. (1989). Role of reticuloendotheliosis virus envelope glycoprotein in superinfection interference. J Virol *63*, 273-280.

DeStefano, J.J., Mallaber, L.M., Rodriguez-Rodriguez, L., Fay, P.J., and Bambara, R.A. (1992). Requirements for strand transfer between internal regions of heteropolymer templates by human immunodeficiency virus reverse transcriptase. J Virol *66*, 6370-6378.

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

Dewannieux, M., and Heidmann, T. (2013). Endogenous retroviruses: acquisition, amplification and taming of genome invaders. Current opinion in virology *3*, 646-656.

Diehl, W.E., Johnson, W.E., and Hunter, E. (2013). Elevated rate of fixation of endogenous retroviral elements in Haplorhini TRIM5 and TRIM22 genomic sequences: impact on transcriptional regulation. PLoS One *8*, e58532.

Diehl, W.E., Patel, N., Halm, K., and Johnson, W.E. (2016). Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals. eLife *5*.

Domazet-Loso, T., and Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. Genome Res *13*, 2213-2219.

Duch, M., Carrasco, M.L., Jespersen, T., Aagaard, L., and Pedersen, F.S. (2004). An RNA secondary structure bias for non-homologous reverse transcriptase-mediated deletions in vivo. Nucleic Acids Res *32*, 2039-2048.

Duckert, P., Brunak, S., and Blom, N. (2004). Prediction of proprotein convertase cleavage sites. Protein Eng Des Sel *17*, 107-112.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

Edwards, R.A., and Rohwer, F. (2005). Viral metagenomics. Nat Rev Microbiol *3*, 504-510.

Emerman, M., and Malik, H.S. (2010). Paleovirology--modern consequences of ancient viruses. PLoS Biol *8*, e1000301.

Esnault, C., Maestre, J., and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. Nat Genet *24*, 363-367.

Fawcett, D.W., and Jensh, R.P. (1997). Bloom & Fawcett : concise histology (New York: Chapman and Hall : International Thomson Pub.).

Feng, Y., Broder, C.C., Kennedy, P.E., and Berger, E.A. (1996). HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. Science *272*, 872-877.

Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. Nature reviews Genetics *13*, 283-296.

Fields, B.N., Griffin, D.E., Howley, P.M., and Knipe, D.M. (2001). Fields' virology, 4th edn (Philadelphia: Lippincott Williams & Wilkins).

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic acids research *39*, W29-37.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S.*, et al.* (2013). Ensembl 2013. Nucleic Acids Res *41*, D48-55.

Fort, A., Hashimoto, K., Yamada, D., Salimullah, M., Keya, C.A., Saxena, A., Bonetti, A., Voineagu, I., Bertin, N., Kratz, A.*, et al.* (2014). Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. Nat Genet *46*, 558-566.

Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. Virus Res *117*, 5-16.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. Nucleic Acids Res *32*, W273-279.

Frost, R.J., Hamra, F.K., Richardson, J.A., Qi, X., Bassel-Duby, R., and Olson, E.N. (2010). MOV10L1 is necessary for protection of spermatocytes against retrotransposons by Piwi-interacting RNAs. Proc Natl Acad Sci U S A *107*, 11847-11852.

Galao, R.P., Le Tortorec, A., Pickering, S., Kueck, T., and Neil, S.J. (2012). Innate sensing of HIV-1 assembly by Tetherin induces NFkappaB-dependent proinflammatory responses. Cell Host Microbe *12*, 633-644.

Gallois-Montbrun, S., Kramer, B., Swanson, C.M., Byers, H., Lynham, S., Ward, M., and Malim, M.H. (2007). Antiviral protein APOBEC3G localizes to ribonucleoprotein complexes found in P bodies and stress granules. J Virol *81*, 2165-2178.

Ganser-Pornillos, B.K., Chandrasekaran, V., Pornillos, O., Sodroski, J.G., Sundquist, W.I., and Yeager, M. (2011). Hexagonal assembly of a restricting TRIM5alpha protein. Proc Natl Acad Sci U S A *108*, 534-539.

Ganser-Pornillos, B.K., Yeager, M., and Sundquist, W.I. (2008). The structural biology of HIV assembly. Curr Opin Struct Biol *18*, 203-217.

Gardner, M.B., Kozak, C.A., and O'Brien, S.J. (1991). The Lake Casitas wild mouse: evolving genetic resistance to retroviral disease. Trends in genetics : TIG *7*, 22-27.

Gifford, R.J., Katzourakis, A., Tristem, M., Pybus, O.G., Winters, M., and Shafer, R.W. (2008). A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. Proc Natl Acad Sci U S A *105*, 20362-20367.

Gilbert, C., Maxfield, D.G., Goodman, S.M., and Feschotte, C. (2009). Parallel germline infiltration of a lentivirus in two Malagasy lemurs. PLoS Genet *5*, e1000425.

Gillick, K., Pollpeter, D., Phalora, P., Kim, E.Y., Wolinsky, S.M., and Malim, M.H. (2013). Suppression of HIV-1 infection by APOBEC3 proteins in primary human CD4(+) T cells is associated with inhibition of processive reverse transcription as well as excessive cytidine deamination. J Virol *87*, 1508-1517.

Goila-Gaur, R., and Strebel, K. (2008). HIV-1 Vif, APOBEC, and intrinsic immunity. Retrovirology *5*, 51.

Goldschmidt, V., Ciuffi, A., Ortiz, M., Brawand, D., Munoz, M., Kaessmann, H., and Telenti, A. (2008). Antiretroviral activity of ancestral TRIM5alpha. J Virol *82*, 2089-2096.

Goldstone, D.C., Yap, M.W., Robertson, L.E., Haire, L.F., Taylor, W.R., Katzourakis, A., Stoye, J.P., and Taylor, I.A. (2010). Structural and functional analysis of prehistoric lentiviruses uncovers an ancient molecular interface. Cell Host Microbe *8*, 248-259.

Goodchild, N.L., Freeman, J.D., and Mager, D.L. (1995). Spliced HERV-H endogenous retroviral sequences in human genomic DNA: evidence for amplification via retrotransposition. Virology *206*, 164-173.

Goujon, C., Moncorge, O., Bauby, H., Doyle, T., Ward, C.C., Schaller, T., Hue, S., Barclay, W.S., Schulz, R., and Malim, M.H. (2013). Human MX2 is an interferon-induced post-entry inhibitor of HIV-1 infection. Nature *502*, 559-562.

Guallar, D., Perez-Palacios, R., Climent, M., Martinez-Abadia, I., Larraga, A., Fernandez-Juan, M., Vallejo, C., Muniesa, P., and Schoorlemmer, J. (2012). Expression of endogenous retroviruses is negatively regulated by the pluripotency marker Rex1/Zfp42. Nucleic Acids Res *40*, 8993-9007.

Gupta, R.K., Hue, S., Schaller, T., Verschoor, E., Pillay, D., and Towers, G.J. (2009). Mutation of a single residue renders human tetherin resistant to HIV-1 Vpu-mediated depletion. PLoS Pathog *5*, e1000443.

Halestrap, A.P., and Wilson, M.C. (2012). The monocarboxylate transporter family--role and regulation. IUBMB Life *64*, 109-119.

Hallenberger, S., Bosch, V., Angliker, H., Shaw, E., Klenk, H.D., and Garten, W. (1992). Inhibition of furin-mediated cleavage activation of HIV-1 glycoprotein gp160. Nature *360*, 358-361.

Han, G.Z., and Worobey, M. (2012). Endogenous lentiviral elements in the weasel family (Mustelidae). Mol Biol Evol *29*, 2905-2908.

Han, G.Z., and Worobey, M. (2015). A primitive endogenous lentivirus in a colugo: insights into the early evolution of lentiviruses. Mol Biol Evol *32*, 211-215.

Hatziioannou, T., Cowan, S., and Bieniasz, P.D. (2004). Capsid-dependent and -independent postentry restriction of primate lentivirus tropism in rodent cells. J Virol *78*, 1006-1011.

Hayward, A., Grabherr, M., and Jern, P. (2013). Broad-scale phylogenomics provides insights into retrovirus-host evolution. Proc Natl Acad Sci U S A *110*, 20146-20151.

Henzy, J.E., and Johnson, W.E. (2013). Pushing the endogenous envelope. Philos Trans R Soc Lond B Biol Sci *368*, 20120506.

Heusinger, E., Kluge, S.F., Kirchhoff, F., and Sauter, D. (2015). Early Vertebrate Evolution of the Host Restriction Factor Tetherin. Journal of virology *89*, 12154-12165.

Hilditch, L., Matadeen, R., Goldstone, D.C., Rosenthal, P.B., Taylor, I.A., and Stoye, J.P. (2011). Ordered assembly of murine leukemia virus capsid protein on lipid nanotubes directs specific binding by the restriction factor, Fv1. Proc Natl Acad Sci U S A *108*, 5771-5776.

Ho, S.H., Shek, L., Gettie, A., Blanchard, J., and Cheng-Mayer, C. (2005). V3 loop-determined coreceptor preference dictates the dynamics of CD4+-T-cell loss in simian-human immunodeficiency virus-infected macaques. J Virol *79*, 12296-12303.

Holmes, E.C. (2011). What does virus evolution tell us about virus origins? J Virol *85*, 5247-5251.

Holmes, R.K., Koning, F.A., Bishop, K.N., and Malim, M.H. (2007). APOBEC3F can inhibit the accumulation of HIV-1 reverse transcription products in the absence of hypermutation. Comparisons with APOBEC3G. The Journal of biological chemistry *282*, 2587-2595.

Holzschu, D.L., Martineau, D., Fodor, S.K., Vogt, V.M., Bowser, P.R., and Casey, J.W. (1995). Nucleotide sequence and protein analysis of a complex piscine retrovirus, walleye dermal sarcoma virus. J Virol *69*, 5320-5331.

Hrecka, K., Hao, C., Gierszewska, M., Swanson, S.K., Kesik-Brodacka, M., Srivastava, S., Florens, L., Washburn, M.P., and Skowronski, J. (2011). Vpx relieves inhibition of HIV-1 infection of macrophages mediated by the SAMHD1 protein. Nature *474*, 658-661.

Hron, T., Fabryova, H., Paces, J., and Elleder, D. (2014). Endogenous lentivirus in Malayan colugo (Galeopterus variegatus), a close relative of primates. Retrovirology *11*, 84.

Hu, W.S., and Hughes, S.H. (2012). HIV-1 reverse transcription. Cold Spring Harb Perspect Med *2*.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T.*, et al.* (2002). The Ensembl genome database project. Nucleic Acids Res *30*, 38-41.

Hulme, A.E., Perez, O., and Hope, T.J. (2011). Complementary assays reveal a relationship between HIV-1 uncoating and reverse transcription. Proc Natl Acad Sci U S A *108*, 9975-9980.

Hwang, S.S., Boyle, T.J., Lyerly, H.K., and Cullen, B.R. (1991). Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. Science *253*, 71-74.

Inaguma, Y., Miyashita, N., Moriwaki, K., Huai, W.C., Jin, M.L., He, X.Q., and Ikeda, H. (1991). Acquisition of two endogenous ecotropic murine leukemia viruses in distinct Asian wild mouse populations. J Virol *65*, 1796-1802.

Inoue, J.G., Miya, M., Lam, K., Tay, B.H., Danks, J.A., Bell, J., Walker, T.I., and Venkatesh, B. (2010). Evolutionary origin and phylogeny of the modern holocephalans (Chondrichthyes: Chimaeriformes): a mitogenomic perspective. Mol Biol Evol *27*, 2576-2586.

International Committee on Taxonomy of Viruses., King, A.M.Q., International Union of Microbiological Societies. Virology Division., and ebrary Inc. (2012). Virus taxonomy classification and nomenclature of viruses : ninth report of the International Committee on Taxonomy of Viruses (London: Elsevier,), pp. x, 1327 p.

Ito, J., Watanabe, S., Hiratsuka, T., Kuse, K., Odahara, Y., Ochi, H., Kawamura, M., and Nishigaki, K. (2013). Refrex-1, a soluble restriction factor against feline endogenous and exogenous retroviruses. J Virol *87*, 12029-12040.

Janvier, P. (2006). Palaeontology: modern look for ancient lamprey. Nature *443*, 921-924.

Jern, P., Sperber, G.O., Ahlsen, G., and Blomberg, J. (2005). Sequence variability, gene structure, and expression of full-length human endogenous retrovirus H. J Virol *79*, 6325-6337.

Ji, X., and Zhao, S. (2008). DA and Xiao-two giant and composite LTR-retrotransposon-like elements identified in the human genome. Genomics *91*, 249-258.

Jia, B., Serra-Moreno, R., Neidermyer, W., Rahmberg, A., Mackey, J., Fofana, I.B., Johnson, W.E., Westmoreland, S., and Evans, D.T. (2009). Species-specific activity of SIV Nef and HIV-1 Vpu in overcoming restriction by tetherin/BST2. PLoS Pathog *5*, e1000429.

Johnson, W.E. (2015). Endogenous Retroviruses in the Genomics Era. Annu Rev Virol *2*, 135-159.

Jolicoeur, P., and Rassart, E. (1980). Effect of Fv-1 gene product on synthesis of linear and supercoiled viral DNA in cells infected with murine leukemia virus. J Virol *33*, 183-195.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res *110*, 462-467.

Kaiser, S.M., Malik, H.S., and Emerman, M. (2007). Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein. Science *316*, 1756-1758.

Kane, M., Yadav, S.S., Bitzegeio, J., Kutluay, S.B., Zang, T., Wilson, S.J., Schoggins, J.W., Rice, C.M., Yamashita, M., Hatziioannou, T.*, et al.* (2013). MX2 is an interferon-induced inhibitor of HIV-1 infection. Nature *502*, 563-566.

Katz, R.A., and Skalka, A.M. (1990). Generation of diversity in retroviruses. Annu Rev Genet *24*, 409-445.

Katzourakis, A., and Gifford, R.J. (2010). Endogenous viral elements in animal genomes. PLoS Genet *6*, e1001191.

Katzourakis, A., Tristem, M., Pybus, O.G., and Gifford, R.J. (2007). Discovery and analysis of the first endogenous lentivirus. Proc Natl Acad Sci U S A *104*, 6261-6265.

Keckesova, Z., Ylinen, L.M., Towers, G.J., Gifford, R.J., and Katzourakis, A. (2009). Identification of a RELIK orthologue in the European hare (Lepus europaeus) reveals a minimum age of 12 million years for the lagomorph lentiviruses. Virology *384*, 7-11.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., and Bosch, T.C. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? Trends in genetics : TIG *25*, 404-413.

Kigami, D., Minami, N., Takayama, H., and Imai, H. (2003). MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. Biol Reprod *68*, 651-654.

Klatzmann, D., Champagne, E., Chamaret, S., Gruest, J., Guetard, D., Hercend, T., Gluckman, J.C., and Montagnier, L. (1984). T-lymphocyte T4 molecule behaves as the receptor for human retrovirus LAV. Nature *312*, 767-768.

Kluge, S.F., Sauter, D., and Kirchhoff, F. (2015). SnapShot: antiviral restriction factors. Cell *163*, 774-774 e771.

Koonin, E.V., Dolja, V.V., and Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology *479-480*, 2-25.

Koonin, E.V., Senkevich, T.G., and Dolja, V.V. (2006). The ancient Virus World and evolution of cells. Biol Direct *1*, 29.

Kosiol, C., Vinar, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R., and Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. PLoS Genet *4*, e1000144.

Kozak, C.A. (2015). Origins of the endogenous and infectious laboratory mouse gammaretroviruses. Viruses *7*, 1-26.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology *305*, 567-580.

Kutluay, S.B., Zang, T., Blanco-Melo, D., Powell, C., Jannain, D., Errando, M., and Bieniasz, P.D. (2014). Global changes in the RNA binding specificity of HIV-1 gag regulate virion genesis. Cell *159*, 1096-1109.

Kvaratskhelia, M., Sharma, A., Larue, R.C., Serrao, E., and Engelman, A. (2014). Molecular mechanisms of retroviral integration site selection. Nucleic Acids Res *42*, 10209-10225.

Laguette, N., Sobhian, B., Casartelli, N., Ringeard, M., Chable-Bessia, C., Segeral, E., Yatim, A., Emiliani, S., Schwartz, O., and Benkirane, M. (2011). SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. Nature *474*, 654-657.

Lahouassa, H., Daddacha, W., Hofmann, H., Ayinde, D., Logue, E.C., Dragin, L., Bloch, N., Maudet, C., Bertrand, M., Gramberg, T.*, et al.* (2012). SAMHD1 restricts the replication of human immunodeficiency virus type 1 by depleting the intracellular pool of deoxynucleoside triphosphates. Nat Immunol *13*, 223-228.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lavialle, C., Cornelis, G., Dupressoir, A., Esnault, C., Heidmann, O., Vernochet, C., and Heidmann, T. (2013). Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. Philos Trans R Soc Lond B Biol Sci *368*, 20120507.

Le Tortorec, A., and Neil, S.J. (2009). Antagonism to and intracellular sequestration of human tetherin by the human immunodeficiency virus type 2 envelope glycoprotein. J Virol *83*, 11966-11978.

Lebedev, Y.B., Belonovitch, O.S., Zybrova, N.V., Khil, P.P., Kurdyukov, S.G., Vinogradova, T.V., Hunsmann, G., and Sverdlov, E.D. (2000). Differences in HERV-K LTR insertions in orthologous loci of humans and great apes. Gene *247*, 265-277.

Lee, A., Nolan, A., Watson, J., and Tristem, M. (2013). Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. Philos Trans R Soc Lond B Biol Sci *368*, 20120503.

Lee, Y.N., and Bieniasz, P.D. (2007). Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog *3*, e10.

Lee, Y.N., Malim, M.H., and Bieniasz, P.D. (2008). Hypermutation of an ancient human retrovirus by APOBEC3G. J Virol *82*, 8762-8770.

Legati, A., Giovannini, D., Nicolas, G., Lopez-Sanchez, U., Quintans, B., Oliveira, J.R., Sears, R.L., Ramos, E.M., Spiteri, E., Sobrido, M.J.*, et al.* (2015). Mutations in XPR1 cause primary familial brain calcification associated with altered phosphate export. Nat Genet *47*, 579-581.

Legrain, P., and Rosbash, M. (1989). Some cis- and trans-acting mutants for splicing target pre-mRNA to the cytoplasm. Cell *57*, 573-583.

Lerner, D.L., Wagaman, P.C., Phillips, T.R., Prospero-Garcia, O., Henriksen, S.J., Fox, H.S., Bloom, F.E., and Elder, J.H. (1995). Increased mutation frequency of feline immunodeficiency virus lacking functional deoxyuridine-triphosphatase. Proc Natl Acad Sci U S A *92*, 7480-7484.

Leung, D.C., and Lorincz, M.C. (2012). Silencing of endogenous retroviruses: when and why do histone marks predominate? Trends Biochem Sci *37*, 127-133.

Lewinski, M.K., Yamashita, M., Emerman, M., Ciuffi, A., Marshall, H., Crawford, G., Collins, F., Shinn, P., Leipzig, J., Hannenhalli, S.*, et al.* (2006). Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS Pathog *2*, e60.

Li, X., Slife, J., Patel, N., and Zhao, S. (2009). Stepwise evolution of two giant composite LTR-retrotransposon-like elements DA and Xiao. BMC Evol Biol *9*, 128.

Liberatore, R.A., and Bieniasz, P.D. (2011). Tetherin is a key effector of the antiretroviral activity of type I interferon in vitro and in vivo. Proc Natl Acad Sci U S A *108*, 18097-18101.

Lilly, F. (1970). Fv-2: identification and location of a second gene governing the spleen focus response to Friend leukemia virus in mice. J Natl Cancer Inst *45*, 163-169.

Lim, E.S., Fregoso, O.I., McCoy, C.O., Matsen, F.A., Malik, H.S., and Emerman, M. (2012). The ability of primate lentiviruses to degrade the monocyte restriction

factor SAMHD1 preceded the birth of the viral accessory protein Vpx. Cell Host Microbe *11*, 194-204.

Liu, Z., Pan, Q., Ding, S., Qian, J., Xu, F., Zhou, J., Cen, S., Guo, F., and Liang, C. (2013). The interferon-inducible MxB protein inhibits HIV-1 infection. Cell Host Microbe *14*, 398-410.

Llorens, C., Fares, M.A., and Moya, A. (2008). Relationships of gag-pol diversity between Ty3/Gypsy and Retroviridae LTR retroelements and the three kings hypothesis. BMC Evol Biol *8*, 276.

Locke, D.P., Hillier, L.W., Warren, W.C., Worley, K.C., Nazareth, L.V., Muzny, D.M., Yang, S.P., Wang, Z., Chinwalla, A.T., Minx, P.*, et al.* (2011). Comparative and demographic analysis of orang-utan genomes. Nature *469*, 529-533.

Low, A., Datta, S., Kuznetsov, Y., Jahid, S., Kothari, N., McPherson, A., and Fan, H. (2007). Mutation in the glycosylated gag protein of murine leukemia virus results in reduced in vivo infectivity and a novel defect in viral budding or release. J Virol *81*, 3685-3692.

Lu, X., Sachs, F., Ramsay, L., Jacques, P.E., Goke, J., Bourque, G., and Ng, H.H. (2014). The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat Struct Mol Biol *21*, 423-425.

Lund, A.H., Duch, M., Lovmand, J., Jorgensen, P., and Pedersen, F.S. (1993). Mutated primer binding sites interacting with different tRNAs allow efficient murine leukemia virus replication. J Virol *67*, 7125-7130.

Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. Science *252*, 1162-1164.

Macfarlan, T.S., Gifford, W.D., Agarwal, S., Driscoll, S., Lettieri, K., Wang, J., Andrews, S.E., Franco, L., Rosenfeld, M.G., Ren, B.*, et al.* (2011). Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes Dev *25*, 594-607.

Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S.L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature *487*, 57-63.

MacMillan, A.L., Kohli, R.M., and Ross, S.R. (2013). APOBEC3 inhibition of mouse mammary tumor virus infection: the role of cytidine deamination versus inhibition of reverse transcription. J Virol *87*, 4808-4817.

Maddon, P.J., Dalgleish, A.G., McDougal, J.S., Clapham, P.R., Weiss, R.A., and Axel, R. (1986). The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain. Cell *47*, 333-348.

Mager, D.L., and Stoye, J.P. (2015). Mammalian Endogenous Retroviruses. Microbiol Spectr *3*, MDNA3-0009-2014.

Magiorkinis, G., Belshaw, R., and Katzourakis, A. (2013). 'There and back again': revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era. Philos Trans R Soc Lond B Biol Sci *368*, 20120504.

Magiorkinis, G., Gifford, R.J., Katzourakis, A., De Ranter, J., and Belshaw, R. (2012). Env-less endogenous retroviruses are genomic superspreaders. Proc Natl Acad Sci U S A *109*, 7385-7390.

Maksakova, I.A., Mager, D.L., and Reiss, D. (2008). Keeping active endogenous retroviral-like elements in check: the epigenetic perspective. Cell Mol Life Sci *65*, 3329-3347.

Malim, M.H., and Bieniasz, P.D. (2012). HIV Restriction Factors and Mechanisms of Evasion. Cold Spring Harb Perspect Med *2*, a006940.

Marin, M., Tailor, C.S., Nouri, A., Kozak, S.L., and Kabat, D. (1999). Polymorphisms of the cell surface receptor control mouse susceptibilities to xenotropic and polytropic leukemia viruses. J Virol *73*, 9362-9368.

Marshall, H.M., Ronen, K., Berry, C., Llano, M., Sutherland, H., Saenz, D., Bickmore, W., Poeschla, E., and Bushman, F.D. (2007). Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. PLoS One *2*, e1340.

Masel, J., and Trotter, M.V. (2010). Robustness and evolvability. Trends in genetics : TIG *26*, 406-414.

Matano, T., Odawara, T., Ohshima, M., Yoshikura, H., and Iwamoto, A. (1993). trans-dominant interference with virus infection at two different stages by a mutant envelope protein of Friend murine leukemia virus. J Virol *67*, 2026-2033.

Matreyek, K.A., and Engelman, A. (2013). Viral and cellular requirements for the nuclear entry of retroviral preintegration nucleoprotein complexes. Viruses *5*, 2483-2511.

McCarthy, K.R., Kirmaier, A., Autissier, P., and Johnson, W.E. (2015). Evolutionary and Functional Analysis of Old World Primate TRIM5 Reveals the Ancient Emergence of Primate Lentiviruses and Convergent Evolution Targeting a Conserved Capsid Interface. PLoS Pathog *11*, e1005085.

McDonald, D., Vodicka, M.A., Lucero, G., Svitkina, T.M., Borisy, G.G., Emerman, M., and Hope, T.J. (2002). Visualization of the intracellular behavior of HIV in living cells. The Journal of cell biology *159*, 441-452.

McNatt, M.W., Zang, T., and Bieniasz, P.D. (2013). Vpu binds directly to tetherin and displaces it from nascent virions. PLoS Pathog *9*, e1003299.

McNatt, M.W., Zang, T., Hatziioannou, T., Bartlett, M., Fofana, I.B., Johnson, W.E., Neil, S.J., and Bieniasz, P.D. (2009). Species-specific activity of HIV-1 Vpu and positive selection of tetherin transmembrane domain variants. PLoS Pathog *5*, e1000300.

Mertz, J.A., Lozano, M.M., and Dudley, J.P. (2009). Rev and Rex proteins of human complex retroviruses function with the MMTV Rem-responsive element. Retrovirology *6*, 10.

Meyerson, N.R., and Sawyer, S.L. (2011). Two-stepping through time: mammals and viruses. Trends Microbiol *19*, 286-294.

Miller, D.G., Edwards, R.H., and Miller, A.D. (1994). Cloning of the cellular receptor for amphotropic murine retroviruses reveals homology to that for gibbon ape leukemia virus. Proc Natl Acad Sci U S A *91*, 78-82.

Moran, J.V., DeBerardinis, R.J., and Kazazian, H.H., Jr. (1999). Exon shuffling by L1 retrotransposition. Science *283*, 1530-1534.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., and Kazazian, H.H., Jr. (1996). High frequency retrotransposition in cultured mammalian cells. Cell *87*, 917-927.

Mouse Genome Sequencing, C., Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M.*, et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Mural, R.J., Adams, M.D., Myers, E.W., Smith, H.O., Miklos, G.L., Wides, R., Halpern, A., Li, P.W., Sutton, G.G., Nadeau, J.*, et al.* (2002). A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science *296*, 1661-1671.

Murcia, P.R., Arnaud, F., and Palmarini, M. (2007). The transdominant endogenous retrovirus enJS56A1 associates with and blocks intracellular trafficking of Jaagsiekte sheep retrovirus Gag. J Virol *81*, 1762-1772.

Neil, S.J., Zang, T., and Bieniasz, P.D. (2008). Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. Nature *451*, 425-430.

Nethe, M., Berkhout, B., and van der Kuyl, A.C. (2005). Retroviral superinfection resistance. Retrovirology *2*, 52.

Nisole, S., Lynch, C., Stoye, J.P., and Yap, M.W. (2004). A Trim5-cyclophilin A fusion protein found in owl monkey kidney cells can restrict HIV-1. Proc Natl Acad Sci U S A *101*, 13324-13328.

Nitta, T., Kuznetsov, Y., McPherson, A., and Fan, H. (2010). Murine leukemia virus glycosylated Gag (gPr80gag) facilitates interferon-sensitive virus release through lipid rafts. Proc Natl Acad Sci U S A *107*, 1190-1195.

Niwa, H., Yamamura, K., and Miyazaki, J. (1991). Efficient selection for high-expression transfectants with a novel eukaryotic vector. Gene *108*, 193-199.

Odaka, T., Ikeda, H., and Akatsuka, T. (1980). Restricted expression of endogenous N-tropic XC-positive leukemia virus in hybrids between G and AKR mice: an effect of the Fv-4r gene. Int J Cancer *25*, 757-762.

Ohlmann, T., Lopez-Lastra, M., and Darlix, J.L. (2000). An internal ribosome entry segment promotes translation of the simian immunodeficiency virus genomic RNA. The Journal of biological chemistry *275*, 11899-11906.

Ohno, S. (1970). Evolution by gene duplication (New York: Springer-Verlag).

Ono, A. (2010). Relationships between plasma membrane microdomains and HIV-1 assembly. Biol Cell *102*, 335-350.

Overbaugh, J., Miller, A.D., and Eiden, M.V. (2001). Receptors and entry cofactors for retroviruses include single and multiple transmembrane-spanning proteins as well as newly described glycophosphatidylinositol-anchored and secreted proteins. Microbiol Mol Biol Rev *65*, 371-389, table of contents.

Padilla-Parra, S., Marin, M., Kondo, N., and Melikyan, G.B. (2014). Pinpointing retrovirus entry sites in cells expressing alternatively spliced receptor isoforms by single virus imaging. Retrovirology *11*, 47.

Palmieri, N., Kosiol, C., and Schlotterer, C. (2014). The life cycle of Drosophila orphan genes. eLife *3*, e01311.

Pan, D., and Zhang, L. (2008). Tandemly arrayed genes in vertebrate genomes. Comparative and functional genomics, 545269.

Pavlicek, A., Paces, J., Elleder, D., and Hejnar, J. (2002). Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution. Genome Res *12*, 391-399.

Pedersen, F.S., Pyrz, M., and Duch, M. (2011). Retroviral Replication. In Encyclopedia of Life Sciences (Chichester, EN: John Wiley & Sons, Ltd.).

Perelman, P., Johnson, W.E., Roos, C., Seuanez, H.N., Horvath, J.E., Moreira, M.A., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y.*, et al.* (2011). A molecular phylogeny of living primates. PLoS Genet *7*, e1001342.

Perez-Caballero, D., Soll, S.J., and Bieniasz, P.D. (2008). Evidence for restriction of ancient primate gammaretroviruses by APOBEC3 but not TRIM5alpha proteins. PLoS Pathog *4*, e1000181.

Perez-Caballero, D., Zang, T., Ebrahimi, A., McNatt, M.W., Gregory, D.A., Johnson, M.C., and Bieniasz, P.D. (2009). Tetherin inhibits HIV-1 release by directly tethering virions to cells. Cell *139*, 499-511.

Pertel, T., Hausmann, S., Morger, D., Zuger, S., Guerra, J., Lascano, J., Reinhard, C., Santoni, F.A., Uchil, P.D., Chatel, L.*, et al.* (2011). TRIM5 is an innate immune sensor for the retrovirus capsid lattice. Nature *472*, 361-365.

Pettit, S.C., Sheng, N., Tritch, R., Erickson-Viitanen, S., and Swanstrom, R. (1998). The regulation of sequential processing of HIV-1 Gag by the viral protease. Adv Exp Med Biol *436*, 15-25.

Picard-Maureau, M., Jarmy, G., Berg, A., Rethwilm, A., and Lindemann, D. (2003). Foamy virus envelope glycoprotein-mediated entry involves a pH-dependent fusion process. J Virol *77*, 4722-4730.

Pierleoni, A., Martelli, P.L., and Casadio, R. (2008). PredGPI: a GPI-anchor predictor. BMC Bioinformatics *9*, 392.

Pillemer, E.A., Kooistra, D.A., Witte, O.N., and Weissman, I.L. (1986). Monoclonal antibody to the amino-terminal L sequence of murine leukemia virus glycosylated gag polyproteins demonstrates their unusual orientation in the cell membrane. J Virol *57*, 413-421.

Pizzato, M. (2010). MLV glycosylated-Gag is an infectivity factor that rescues Nef-deficient HIV-1. Proc Natl Acad Sci U S A *107*, 9364-9369.

Prats, A.C., De Billy, G., Wang, P., and Darlix, J.L. (1989). CUG initiation codon used for the synthesis of a cell surface antigen coded by the murine leukemia virus. Journal of molecular biology *205*, 363-372.

Prufer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J.R., Walenz, B., Koren, S., Sutton, G., Kodira, C., Winer, R.*, et al.* (2012). The bonobo genome compared with the chimpanzee and human genomes. Nature *486*, 527-531.

Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., *et al.* (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res *42*, D756-763.

Rai, S.K., Duh, F.M., Vigdorovich, V., Danilkovitch-Miagkova, A., Lerman, M.I., and Miller, A.D. (2001). Candidate tumor suppressor HYAL2 is a glycosylphosphatidylinositol (GPI)-anchored cell-surface receptor for jaagsiekte sheep retrovirus, the envelope protein of which mediates oncogenic transformation. Proc Natl Acad Sci U S A *98*, 4443-4448.

Rambaut, A. (2008). FigTree v1.4.2: Tree figure drawing tool. Available at http://treebioedacuk/software/figtree/.

Rambaut, A., and Grassly, N.C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci *13*, 235-238.

Rebollo, R., Romanish, M.T., and Mager, D.L. (2012). Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet *46*, 21-42.

Rehwinkel, J. (2014). Mouse knockout models for HIV-1 restriction factors. Cell Mol Life Sci *71*, 3749-3766.

Rein, A. (2010). Nucleic acid chaperone activity of retroviral Gag proteins. RNA Biol *7*, 700-705.

Rein, A., Mirro, J., Haynes, J.G., Ernst, S.M., and Nagashima, K. (1994). Function of the cytoplasmic domain of a retroviral transmembrane protein: p15E-p2E cleavage activates the membrane fusion capability of the murine leukemia virus Env protein. J Virol *68*, 1773-1781.

Repaske, R., Steele, P.E., O'Neill, R.R., Rabson, A.B., and Martin, M.A. (1985). Nucleotide sequence of a full-length human endogenous retroviral segment. J Virol *54*, 764-772.

Rhesus Macaque Genome, S., Analysis, C., Gibbs, R.A., Rogers, J., Katze, M.G., Bumgarner, R., Weinstock, G.M., Mardis, E.R., Remington, K.A., Strausberg, R.L., *et al.* (2007). Evolutionary and biomedical insights from the rhesus macaque genome. Science *316*, 222-234.

Ribet, D., Harper, F., Dupressoir, A., Dewannieux, M., Pierron, G., and Heidmann, T. (2008a). An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus. Genome Res *18*, 597-609.

Ribet, D., Louvet-Vallee, S., Harper, F., de Parseval, N., Dewannieux, M., Heidmann, O., Pierron, G., Maro, B., and Heidmann, T. (2008b). Murine endogenous retrovirus MuERV-L is the progenitor of the "orphan" epsilon viruslike particles of the early mouse embryo. J Virol *82*, 1622-1625.

Rice, G.I., Bond, J., Asipu, A., Brunette, R.L., Manfield, I.W., Carr, I.M., Fuller, J.C., Jackson, R.M., Lamb, T., Briggs, T.A.*, et al.* (2009). Mutations involved in Aicardi-Goutieres syndrome implicate SAMHD1 as regulator of the innate immune response. Nat Genet *41*, 829-832.

Robinson, H.L., Astrin, S.M., Senior, A.M., and Salazar, F.H. (1981). Host Susceptibility to endogenous viruses: defective, glycoprotein-expressing proviruses interfere with infections. J Virol *40*, 745-751.

Roe, T., Reynolds, T.C., Yu, G., and Brown, P.O. (1993). Integration of murine leukemia virus DNA depends on mitosis. EMBO J *12*, 2099-2108.

Romanish, M.T., Lock, W.M., van de Lagemaat, L.N., Dunn, C.A., and Mager, D.L. (2007). Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. PLoS Genet *3*, e10.

Rosa, A., Chande, A., Ziglio, S., De Sanctis, V., Bertorelli, R., Goh, S.L., McCauley, S.M., Nowosielska, A., Antonarakis, S.E., Luban, J.*, et al.* (2015). HIV-1 Nef promotes infection by excluding SERINC5 from virion incorporation. Nature *526*, 212-217.

Rose, P.P., and Korber, B.T. (2000). Detecting hypermutations in viral sequences with an emphasis on G --> A hypermutation. Bioinformatics *16*, 400-401.

Rowe, H.M., and Trono, D. (2011). Dynamic control of endogenous retroviruses during development. Virology *411*, 273-287.

Ryoo, J., Choi, J., Oh, C., Kim, S., Seo, M., Kim, S.Y., Seo, D., Kim, J., White, T.E., Brandariz-Nunez, A.*, et al.* (2014). The ribonuclease activity of SAMHD1 is required for HIV-1 restriction. Nat Med *20*, 936-941.

Saad, J.S., Miller, J., Tai, J., Kim, A., Ghanam, R.H., and Summers, M.F. (2006). Structural basis for targeting HIV-1 Gag proteins to the plasma membrane for virus assembly. Proc Natl Acad Sci U S A *103*, 11364-11369.

Salzwedel, K., Smith, E.D., Dey, B., and Berger, E.A. (2000). Sequential CD4-coreceptor interactions in human immunodeficiency virus type 1 Env function: soluble CD4 activates Env for coreceptor-dependent fusion and reveals blocking activities of antibodies against cryptic conserved epitopes on gp120. J Virol *74*, 326-333.

Sattentau, Q.J., Clapham, P.R., Weiss, R.A., Beverley, P.C., Montagnier, L., Alhalabi, M.F., Gluckmann, J.C., and Klatzmann, D. (1988). The human and simian immunodeficiency viruses HIV-1, HIV-2 and SIV interact with similar epitopes on their cellular receptor, the CD4 molecule. AIDS *2*, 101-105.

Sattler, S., Reiche, D., Sturtzel, C., Karas, I., Richter, S., Kalb, M.L., Gregor, W., and Hofer, E. (2012). The human C-type lectin-like receptor CLEC-1 is upregulated by TGF-beta and primarily localized in the endoplasmic membrane compartment. Scandinavian journal of immunology *75*, 282-292.

Sawyer, S.L., Emerman, M., and Malik, H.S. (2004). Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS Biol *2*, E275.

Sawyer, S.L., Wu, L.I., Emerman, M., and Malik, H.S. (2005). Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci U S A *102*, 2832-2837.

Sayah, D.M., Sokolskaja, E., Berthoux, L., and Luban, J. (2004). Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature *430*, 569-573.

Scally, A., Dutheil, J.Y., Hillier, L.W., Jordan, G.E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T.*, et al.* (2012). Insights into hominid evolution from the gorilla genome sequence. Nature *483*, 169-175.

Schaller, T., Ylinen, L.M., Webb, B.L., Singh, S., and Towers, G.J. (2007). Fusion of cyclophilin A to Fv1 enables cyclosporine-sensitive restriction of human and feline immunodeficiency viruses. J Virol *81*, 10055-10063.

Schroder, A.R., Shinn, P., Chen, H., Berry, C., Ecker, J.R., and Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. Cell *110*, 521-529.

Serra-Moreno, R., Jia, B., Breed, M., Alvarez, X., and Evans, D.T. (2011). Compensatory changes in the cytoplasmic tail of gp41 confer resistance to tetherin/BST-2 in a pathogenic nef-deleted SIV. Cell Host Microbe *9*, 46-57.

Shaffer, H.B., Minx, P., Warren, D.E., Shedlock, A.M., Thomson, R.C., Valenzuela, N., Abramyan, J., Amemiya, C.T., Badenhorst, D., Biggar, K.K.*, et al.* (2013). The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. Genome Biol *14*, R28.

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

Shioda, T., Levy, J.A., and Cheng-Mayer, C. (1991). Macrophage and T cell-line tropisms of HIV-1 are determined by specific regions of the envelope gp120 gene. Nature *349*, 167-169.

Simmons, G.S., Young, P.R., Hanger, J.J., Jones, K., Clarke, D., McKee, J.J., and Meers, J. (2012). Prevalence of koala retrovirus in geographically diverse populations in Australia. Aust Vet J *90*, 404-409.

Smith, J.J., Kuraku, S., Holt, C., Sauka-Spengler, T., Jiang, N., Campbell, M.S., Yandell, M.D., Manousaki, T., Meyer, A., Bloom, O.E.*, et al.* (2013). Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nat Genet *45*, 415-421, 421e411-412.

Sodroski, J., Patarca, R., Rosen, C., Wong-Staal, F., and Haseltine, W. (1985a). Location of the trans-activating region on the genome of human T-cell lymphotropic virus type III. Science *229*, 74-77.

Sodroski, J., Rosen, C., Wong-Staal, F., Salahuddin, S.Z., Popovic, M., Arya, S., Gallo, R.C., and Haseltine, W.A. (1985b). Trans-acting transcriptional regulation of human T-cell leukemia virus type III long terminal repeat. Science *227*, 171-173.

Soll, S.J., Neil, S.J., and Bieniasz, P.D. (2010). Identification of a receptor for an extinct virus. Proc Natl Acad Sci U S A *107*, 19496-19501.

Sommerfelt, M.A., and Weiss, R.A. (1990). Receptor interference groups of 20 retroviruses plating on human cells. Virology *176*, 58-69.

Soneoka, Y., Cannon, P.M., Ramsdale, E.E., Griffiths, J.C., Romano, G., Kingsman, S.M., and Kingsman, A.J. (1995). A transient three-plasmid expression system for the production of high titer retroviral vectors. Nucleic Acids Res *23*, 628-633.

Song, B., Gold, B., O'Huigin, C., Javanbakht, H., Li, X., Stremlau, M., Winkler, C., Dean, M., and Sodroski, J. (2005). The B30.2(SPRY) domain of the retroviral restriction factor TRIM5alpha exhibits lineage-specific length and sequence variation in primates. J Virol *79*, 6111-6121.

Souvorov, A., Kapustin, Y., Kiryutin, B., Chetvernin, V., Tatusova, T., and Lipman, D. (2010). Gnomon - NCBI eukaryotic gene prediction tool. (http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml.).

Spencer, T.E., Mura, M., Gray, C.A., Griebel, P.J., and Palmarini, M. (2003). Receptor usage and fetal expression of ovine endogenous betaretroviruses: implications for coevolution of endogenous and exogenous retroviruses. J Virol *77*, 749-753.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688-2690.

Stan, R.V., Ghitescu, L., Jacobson, B.S., and Palade, G.E. (1999a). Isolation, cloning, and localization of rat PV-1, a novel endothelial caveolar protein. The Journal of cell biology *145*, 1189-1198.

Stan, R.V., Kubitza, M., and Palade, G.E. (1999b). PV-1 is a component of the fenestral and stomatal diaphragms in fenestrated endothelia. Proc Natl Acad Sci U S A *96*, 13203-13207.

Stan, R.V., Tse, D., Deharvengt, S.J., Smits, N.C., Xu, Y., Luciano, M.R., McGarry, C.L., Buitendijk, M., Nemani, K.V., Elgueta, R.*, et al.* (2012). The diaphragms of fenestrated endothelia: gatekeepers of vascular permeability and blood composition. Dev Cell *23*, 1203-1218.

Stavrou, S., Nitta, T., Kotla, S., Ha, D., Nagashima, K., Rein, A.R., Fan, H., and Ross, S.R. (2013). Murine leukemia virus glycosylated Gag blocks apolipoprotein B editing complex 3 and cytosolic sensor access to the reverse transcription complex. Proc Natl Acad Sci U S A *110*, 9078-9083.

Stein, B.S., and Engleman, E.G. (1990). Intracellular processing of the gp160 HIV-1 envelope precursor. Endoproteolytic cleavage occurs in a cis or medial compartment of the Golgi complex. The Journal of biological chemistry *265*, 2640-2649.

Stetson, D.B., Ko, J.S., Heidmann, T., and Medzhitov, R. (2008). Trex1 prevents cell-intrinsic initiation of autoimmunity. Cell *134*, 587-598.

Stremlau, M., Owens, C.M., Perron, M.J., Kiessling, M., Autissier, P., and Sodroski, J. (2004). The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. Nature *427*, 848-853.

Stremlau, M., Perron, M., Lee, M., Li, Y., Song, B., Javanbakht, H., Diaz-Griffero, F., Anderson, D.J., Sundquist, W.I., and Sodroski, J. (2006). Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5alpha restriction factor. Proc Natl Acad Sci U S A *103*, 5514-5519.

Subramanian, R.P., Wildschutte, J.H., Russo, C., and Coffin, J.M. (2011). Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology *8*, 90.

Sundquist, W.I., and Krausslich, H.G. (2012). HIV-1 assembly, budding, and maturation. Cold Spring Harb Perspect Med *2*, a006924.

Suttle, C.A. (2007). Marine viruses--major players in the global ecosystem. Nat Rev Microbiol *5*, 801-812.

Sverdlov, E.D. (2005). Retroviruses and primate genome evolution (Georgetown, Tex.: Landes Bioscience).

Swofford, D.L. (2002). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4 Sinauer Associates, Sunderland, Mass.

Tailor, C.S., Lavillette, D., Marin, M., and Kabat, D. (2003). Cell surface receptors for gammaretroviruses. Current topics in microbiology and immunology *281*, 29-106.

Tailor, C.S., Nouri, A., Lee, C.G., Kozak, C., and Kabat, D. (1999). Cloning and characterization of a cell surface receptor for xenotropic and polytropic murine leukemia viruses. Proc Natl Acad Sci U S A *96*, 927-932.

Takeda, E., Nakagawa, S., Nakaya, Y., Tanaka, A., Miyazawa, T., and Yasuda, J. (2012). Identification and functional analysis of three isoforms of bovine BST-2. PLoS One *7*, e41483.

Tarlinton, R.E., Meers, J., and Young, P.R. (2006). Retroviral invasion of the koala genome. Nature *442*, 79-81.

Tautz, D., and Domazet-Loso, T. (2011). The evolutionary origin of orphan genes. Nature reviews Genetics *12*, 692-702.

Taylor, G.M., Gao, Y., and Sanders, D.A. (2001). Fv-4: identification of the defect in Env and the mechanism of resistance to ecotropic murine leukemia virus. J Virol *75*, 11244-11248.

Temin, H.M., and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature *226*, 1211-1213.

Thomas, J.H., and Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. Genome Res *21*, 1800-1812.

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

Usami, Y., Popov, S., and Gottlinger, H.G. (2014). The Nef-like effect of murine leukemia virus glycosylated gag on HIV-1 infectivity is mediated by its cytoplasmic domain and depends on the AP-2 adaptor complex. J Virol *88*, 3443-3454.

Usami, Y., Wu, Y., and Gottlinger, H.G. (2015). SERINC3 and SERINC5 restrict HIV-1 infectivity and are counteracted by Nef. Nature *526*, 218-223.

Van Damme, N., Goff, D., Katsura, C., Jorgenson, R.L., Mitchell, R., Johnson, M.C., Stephens, E.B., and Guatelli, J. (2008). The interferon-induced protein BST-2 restricts HIV-1 release and is downregulated from the cell surface by the viral Vpu protein. Cell Host Microbe *3*, 245-252.

van der Kuyl, A.C. (2012). HIV infection and HERV expression: a review. Retrovirology *9*, 6.

van Zeijl, M., Johann, S.V., Closs, E., Cunningham, J., Eddy, R., Shows, T.B., and O'Hara, B. (1994). A human amphotropic retrovirus receptor is a second member of the gibbon ape leukemia virus receptor family. Proc Natl Acad Sci U S A *91*, 1168-1172.

Varela, M., Spencer, T.E., Palmarini, M., and Arnaud, F. (2009). Friendly viruses: the special relationship between endogenous retroviruses and their host. Ann N Y Acad Sci *1178*, 157-172.

Veillette, M., Bichel, K., Pawlica, P., Freund, S.M., Plourde, M.B., Pham, Q.T., Reyes-Moreno, C., James, L.C., and Berthoux, L. (2013). The V86M mutation in HIV-1 capsid confers resistance to TRIM5alpha by abrogation of cyclophilin A-dependent restriction and enhancement of viral nuclear import. Retrovirology *10*, 25.

Venkatesh, B., Lee, A.P., Ravi, V., Maurya, A.K., Lian, M.M., Swann, J.B., Ohta, Y., Flajnik, M.F., Sutoh, Y., Kasahara, M*., et al.* (2014). Elephant shark genome provides unique insights into gnathostome evolution. Nature *505*, 174-179.

Venkatesh, S., and Bieniasz, P.D. (2013). Mechanism of HIV-1 virion entrapment by tetherin. PLoS Pathog *9*, e1003483.

Villesen, P., Aagaard, L., Wiuf, C., and Pedersen, F.S. (2004). Identification of endogenous retroviral reading frames in the human genome. Retrovirology *1*, 32.

Virgen, C.A., Kratovac, Z., Bieniasz, P.D., and Hatziioannou, T. (2008). Independent genesis of chimeric TRIM5-cyclophilin proteins in two primate species. Proc Natl Acad Sci U S A *105*, 3563-3568.

Von Hoegen, I., Nakayama, E., and Parnes, J.R. (1990). Identification of a human protein homologous to the mouse Lyb-2 B cell differentiation antigen and sequence of the corresponding cDNA. Journal of immunology *144*, 4870-4877.

Wagner, A. (2011). The molecular origins of evolutionary innovations. Trends in genetics : TIG *27*, 397-410.

Wang, E., Obeng-Adjei, N., Ying, Q., Meertens, L., Dragic, T., Davey, R.A., and Ross, S.R. (2008). Mouse mammary tumor virus uses mouse but not human transferrin receptor 1 to reach a low pH compartment and infect cells. Virology *381*, 230-240.

Wang, H., Klamo, E., Kuhmann, S.E., Kozak, S.L., Kavanaugh, M.P., and Kabat, D. (1996). Modulation of ecotropic murine retroviruses by N-linked glycosylation of the cell surface receptor/amino acid transporter. J Virol *70*, 6884-6891.

Wang, J., Xie, G., Singh, M., Ghanbarian, A.T., Rasko, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N.V.*, et al.* (2014). Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature *516*, 405-409.

Weiss, R. (1967). Spontaneous virus production from "non-virus producing" Rous sarcoma cells. Virology *32*, 719-723.

Weiss, R.A. (2006). The discovery of endogenous retroviruses. Retrovirology *3*, 67.

Weiss, R.A., and Tailor, C.S. (1995). Retrovirus receptors. Cell *82*, 531-533.

Wildschutte, J.H., Williams, Z.H., Montesion, M., Subramanian, R.P., Kidd, J.M., and Coffin, J.M. (2016). Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc Natl Acad Sci U S A.

Wilson, S.J., Webb, B.L., Ylinen, L.M., Verschoor, E., Heeney, J.L., and Towers, G.J. (2008). Independent evolution of an antiviral TRIMCyp in rhesus macaques. Proc Natl Acad Sci U S A *105*, 3557-3562.

Wu, T., Yan, Y., and Kozak, C.A. (2005). Rmcf2, a xenotropic provirus in the Asian mouse species Mus castaneus, blocks infection by polytropic mouse gammaretroviruses. J Virol *79*, 9677-9684.

Wu, X., Li, Y., Crise, B., and Burgess, S.M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. Science *300*, 1749-1751.

Xie, D., Chen, C.C., Ptaszek, L.M., Xiao, S., Cao, X., Fang, F., Ng, H.H., Lewin, H.A., Cowan, C., and Zhong, S. (2010). Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. Genome Res *20*, 804-815.

Xu, X., Nagarajan, H., Lewis, N.E., Pan, S., Cai, Z., Liu, X., Chen, W., Xie, M., Wang, W., Hammond, S.*, et al.* (2011). The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nat Biotechnol *29*, 735-741.

Yamaguchi, S., Hasegawa, M., Suzuki, T., Ikeda, H., Aizawa, S., Hirokawa, K., and Kitagawa, M. (2003). In vivo distribution of receptor for ecotropic murine leukemia virus and binding of envelope protein of Friend Murine leukemia virus. Arch Virol *148*, 1175-1184.

Yan, Y., Buckler-White, A., Wollenberg, K., and Kozak, C.A. (2009). Origin, antiviral function and evidence for positive selection of the gammaretrovirus restriction gene Fv1 in the genus Mus. Proc Natl Acad Sci U S A *106*, 3259-3263.

Yang, W.K., Kiggans, J.O., Yang, D.M., Ou, C.Y., Tennant, R.W., Brown, A., and Bassin, R.H. (1980). Synthesis and circularization of N- and B-tropic retroviral DNA Fv-1 permissive and restrictive mouse cells. Proc Natl Acad Sci U S A *77*, 2994-2998.

Yang, Y.L., Guo, L., Xu, S., Holland, C.A., Kitamura, T., Hunter, K., and Cunningham, J.M. (1999). Receptors for polytropic and xenotropic mouse leukaemia viruses encoded by a single gene at Rmc1. Nat Genet *21*, 216-219.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci *13*, 555-556.

Yap, M.W., Dodding, M.P., and Stoye, J.P. (2006). Trim-cyclophilin A fusion proteins can restrict human immunodeficiency virus type 1 infection at two distinct phases in the viral life cycle. J Virol *80*, 4061-4067.

Yap, M.W., Nisole, S., Lynch, C., and Stoye, J.P. (2004). Trim5alpha protein restricts both HIV-1 and murine leukemia virus. Proc Natl Acad Sci U S A *101*, 10786-10791.

Yu, Q., Konig, R., Pillai, S., Chiles, K., Kearney, M., Palmer, S., Richman, D., Coffin, J.M., and Landau, N.R. (2004). Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. Nat Struct Mol Biol *11*, 435-442.

Zennou, V., Perez-Caballero, D., Gottlinger, H., and Bieniasz, P.D. (2004). APOBEC3G incorporation into human immunodeficiency virus type 1 particles. J Virol *78*, 12058-12061.

Zhang, F., Hatziioannou, T., Perez-Caballero, D., Derse, D., and Bieniasz, P.D. (2006). Antiretroviral potential of human tripartite motif-5 and related proteins. Virology *353*, 396-409.

Zhang, F., Wilson, S.J., Landford, W.C., Virgen, B., Gregory, D., Johnson, M.C., Munch, J., Kirchhoff, F., Bieniasz, P.D., and Hatziioannou, T. (2009). Nef proteins from simian immunodeficiency viruses are tetherin antagonists. Cell Host Microbe *6*, 54-67.

Zheng, Y.H., Jeang, K.T., and Tokunaga, K. (2012). Host restriction factors in retroviral infection: promises in virus-host interaction. Retrovirology *9*, 112.