Genomic Variation and Evolution of HERV-H and other Endogenous Retroviruses (ERVs)

PATRIC JERN
Dissertation presented at Uppsala University to be publicly examined in Föreläsningssalen, Klinisk virologi, Akademiska Sjukhuset, ing. D1, Uppsala, Saturday, October 1, 2005 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract

An exogenous retrovirus (XRV) that integrates into a germ cell may be inherited as a Mendelian gene; it becomes an endogenous retrovirus (ERV). The human genome consists of up to 8% HERVs.

The gammaretroviral (ERV class I) HERV-H, with 926 members, is the largest ERV group. Despite millions of years since integration, it has polymorphic envelope open reading frames in at least three loci. Selections for functional envelopes are indicated on chromosomes 1 and 2. However, envelopes were present only in a fraction of the total HERV-H. Mutated polymerases, indicating old ERVs, contradicted relatively intact long terminal repeats. To explain this, we formulated a “Midwife” element theory where proteins are complemented in trans.

A phylogenetic analysis did not support separate HERV-H and -F groups. The new taxonomy included HERV-H like (RGH2-like and RTVLH2-like subgroups) and Adjacent HERV-H like. A bioinformatic reconstruction of a putative ancestral HERV-H exposed novel traits. Two nucloecapsid zinc fingers and a pronounced nucleotide bias for C in the HERV-H like were unique among the gammaretroviruses.

Two recently integrated gammaretroviral groups (PtNeo-1[PTERV1]) and -II) were found in chimpanzees but not in humans. The PtNeo groups were most similar to baboon ERVs and a macaque sequence, but neither to other chimpanzee nor to any human gammaretroviruses. The pattern was consistent with cross-species transfer via predation.

To advance the retroviral taxonomy, we projected structural markers over sequence phylogenetic trees. A number of markers were useful to distinguish between genera and to delineate groups.

Basic retroviral knowledge is vital to understand emerging infections. Phylogenetic analysis of taxonomically improved sequences, facilitates the search for common retroviral denominators to target. This thesis provided new insights in retroviral evolution and taxonomy using the ERVs, with special focus on the large gammaretroviral HERV-H group, as an additional source of information next to that of XRVs.

Keywords: Endogenous Retrovirus (ERV), HERV-H, Phylogeny, Evolution, Sequence Variation, Polymorphism, Recent Integrations, Horizontal Transfer, Master Elements, Taxonomy

Patric Jern, Department of Medical Sciences, Akademiska sjukhuset, Uppsala University, SE-75185 Uppsala, Sweden

© Patric Jern 2005

ISSN 1651-6206
ISBN 91-554-6319-3
urn:nbn:se:uu:diva-5906 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-5906)
List of Papers

This thesis is based on following publications, which are referred to in the text by their roman numerals.

I   **Jern P.,** Lindeskog M., Karlsson D., and Blomberg J.
    Full-Length HERV-H Elements with env SU Open Reading Frames in the Human Genome.
    *AIDS Research and Human Retroviruses*, 18(9); 671-676. 2002.¹

II  **Jern P.,** Sperber G.O., and Blomberg J.
    Definition and Variation of Human Endogenous Retrovirus H.
    *Virology*, 327(1); 93-110. 2004.²

III **Jern P.,** Sperber G.O., Ahlsén G., and Blomberg J.
    Sequence Variability, Gene Structure and Expression of Full-Length Human Endogenous Retrovirus H.
    *J. Virology*, 79(10); 6325-6337. 2005.³

IV  **Jern P.,** Sperber G.O., and Blomberg J.
    Divergent Patterns of Recent Retroviral Integration in Human and Chimpanzee Genomes; Transmissions from Other Primates to Chimpanzees.
    *(In progress).*

V   **Jern P.,** Sperber G.O., and Blomberg J.
    Use of Endogenous Retroviral Sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy.
    *Retrovirology*, 2(50). 2005.⁴

**Related work (not included in this thesis)**

VI  **Blomberg J.,** Ushameckis D., and **Jern P.**
    Evolutionary Aspects of Human Endogenous Retroviral Sequences (HERVs) and Disease.
Sperber G.O., Airola T., Jern P., Castell A., and Blomberg J.
RetroTector© Automated Recognition of Retroviral Sequences in Genomic Data.
(Manuscript)

Reprints were made with permissions from the publishers.
1 © 2002, Mary Ann Liebert Inc. (USA).
2 © 2004, Elsevier Inc. (USA).
3 © 2005, American Society For Microbiology, ASM. (USA).
4 © 2005, Jern et al; licensee BioMed Central Ltd.
Contents

Introduction.....................................................................................................9
Background and a short history .................................................................9
Retroelements..........................................................................................10
  Processed pseudogenes ....................................................................12
  Retroposons ..................................................................................12
  Retrotransposons .........................................................................13
  Retroviruses ..................................................................................13
    Retroviral genome structure .....................................................14
    Retroviral replication ................................................................17
Endogenous retroviruses (ERVs) .............................................................20
  Endogenization .............................................................................20
  Biological relevance of ERVs ..........................................................20
    Promoters and enhancers .............................................................20
    Alternative and intergenic splicing .............................................21
    Impact of retroviral Env .............................................................23
  Retroviral restriction ....................................................................24
  Cancer ............................................................................................25
  Neurodegenerative disorders and autoimmunity ..............................27
Retroviral classification and taxonomy ....................................................29
  Morphology based classification ....................................................29
  Sequence based classification (XRV and ERV) .................................30
    Retroviral genera; Pol sequence based classification ....................33
      Alpharetroviruses ..................................................................33
      Betaretroviruses ..................................................................33
      Gammaretroviruses ................................................................34
      Deltaretroviruses ..................................................................34
      Epsilonretroviruses ................................................................35
      Lentiviruses ..........................................................................35
      Spuma and Spuma-like retroviruses .......................................35
      Novel intermediate sequence groups .....................................36
ERV evolution; forces and consequences .................................................36
  Infection and reinfection ................................................................36
  Polymorphism ..............................................................................38
  Recombination and gene conversion ..............................................39
  Horizontal transfer ......................................................................40
Master element theories; ERV “Break out” vs. “Midwife” ...............41

The present investigation ........................................................................................................44
The large HERV-H/F like group ..................................................................................44
Aims of the study .............................................................................................................47
Paper I ..................................................................................................................................47
Paper II .............................................................................................................................48
Paper III ............................................................................................................................49
Paper IV ............................................................................................................................51
Paper V ..................................................................................................................................53

Concluding remarks and future prospects .................................................................56

Acknowledgements ...........................................................................................................58

References ...........................................................................................................................59

Appendices ..........................................................................................................................73
Appendix A. Retroviral (XRV and ERV) nucleotide frequencies ......................74
Appendix B. Putative ancestral HERV-H sequence ..............................................75
Abbreviations

General abbreviations

aa    amino acid
bp    base pairs (nucleotides)
CA    capsid protein
cDNA  complementary deoxy-ribonucleic acid
DNA   deoxy-ribonucleic acid
dUTPase deoxy-uracil triphosphatase
env   envelope gene
gag   group specific antigen gene
IN    integrase domain
kb    kilo basepairs
kDa   kilo Dalton (protein molecular size)
LINE  long interspersed nucleotide element
LTR   long terminal repeat
MA    matrix protein
nt    nucleotide(s)
NC    nucleocapsid protein
ORF   open reading frame
PCR   polymerase chain reaction (nucleic acid primer extention)
PBS   primer binding site
pol   polymerase gene
pro   protease gene
R     repeat sequence (in the LTR)
RH    RNase H domain
RT    reverse transcriptase domain
SA    splice acceptor site
SD    splice donor site
SINE  short interspersed nucleotide element
SU    (envelope) surface unit
TE    transposable element
TM    (envelope) transmembrane protein
tRNA  transfer ribonucleic acid
U3    unique 3’-sequence
U5    unique 5’-sequence
**Retroviral abbreviations**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALV</td>
<td>avian leukosis virus</td>
</tr>
<tr>
<td>BaEV</td>
<td>baboon endogenous retrovirus</td>
</tr>
<tr>
<td>BLV</td>
<td>bovine leukemia virus</td>
</tr>
<tr>
<td>EIAV</td>
<td>equine infectious anemia virus</td>
</tr>
<tr>
<td>ERV</td>
<td>endogenous retrovirus</td>
</tr>
<tr>
<td>FIV</td>
<td>feline immunodeficiency virus</td>
</tr>
<tr>
<td>GaLV</td>
<td>gibbon ape leukemia virus</td>
</tr>
<tr>
<td>HERV</td>
<td>human endogenous retrovirus</td>
</tr>
<tr>
<td>HFV</td>
<td>human foamy virus</td>
</tr>
<tr>
<td>HIV</td>
<td>human immunodeficiency virus</td>
</tr>
<tr>
<td>HML</td>
<td>human mouse mammary tumour virus like</td>
</tr>
<tr>
<td>HSRV</td>
<td>human spuma retrovirus</td>
</tr>
<tr>
<td>HTLV</td>
<td>human t-lymphotropic virus</td>
</tr>
<tr>
<td>JSRV</td>
<td>jaagsiekte (sheep) retrovirus</td>
</tr>
<tr>
<td>MLV</td>
<td>murine leukemia virus</td>
</tr>
<tr>
<td>MMTV</td>
<td>mouse mammary tumour virus</td>
</tr>
<tr>
<td>MoLV</td>
<td>moloney murine leukemia virus</td>
</tr>
<tr>
<td>MPMV</td>
<td>mason-pfiizer monkey virus</td>
</tr>
<tr>
<td>RSV</td>
<td>rous sarcoma virus</td>
</tr>
<tr>
<td>SnRV</td>
<td>snakehead retrovirus</td>
</tr>
<tr>
<td>WDSV</td>
<td>walleye dermal sarcoma virus</td>
</tr>
<tr>
<td>Xen1</td>
<td><em>Xenopus laevis</em> retrovirus 1</td>
</tr>
<tr>
<td>XRV</td>
<td>exogenous retrovirus</td>
</tr>
</tbody>
</table>
Introduction

Background and a short history

A retrovirus has an encapsidated double setup of positive single stranded ribonucleic acids (RNAs), which is enclosed in a lipid bilayer envelope. The retrovirus replicates by conversion from RNA to complementary deoxyribonucleic acid (cDNA) using the virus encoded reverse transcriptase (RT) (Baltimore, 1970; Temin & Mizutani, 1970). The function of RT was in contradiction to the central dogma, proposed by Crick in 1958, and forced an update of the central dogma of molecular biology in 1970 (Crick, 1970). The retroviral integrase (IN) inserts the cDNA into the host genome, where it resides as a provirus (Temin, 1964), and may be expressed to produce progeny.

In the beginning of the 20th century, researchers Ellermann and Bang in 1908 and Rous in 1911 discovered the first retroviruses in chicken suffering from leukosis. A few decades later, in 1936, a mammalian retrovirus associated with mammary carcinoma was detected in mice by Dr. J.J. Bittner (Bittner, 1936). For a historical review see P.K. Vogt (1997a). The study of retroviral disease in domesticated animals has thereafter been encouraged by the viral effects and the possibility to use the animals as model systems for analogues to human diseases. The research on retroviruses increased during the 1970s and resulted, in the 1980s, in the discovery of the causative agents for human T-cell leukemia (by human T-cell lymphotropic virus, HTLV (Poiesz et al., 1980)) and the cause of the acquired immunodeficiency syndrome, AIDS (by human immunodeficiency virus, HIV (Barre-Sinoussi et al., 1983; Gallo et al., 1984)). At that time reports also started to appear about human endogenous retroviruses (HERVs) (Martin et al., 1981; Bonner et al., 1982).

Occasionally, integration of a retrovirus into a germ cell may lead to a provirus that is passed to the offspring and the retrovirus is inherited as a Mendelian gene; it becomes an endogenous retrovirus (ERV) (Stoye & Coffin, 1985). HERVs have been estimated to constitute about 7-8 % of the human
genome (Smit, 1999; Bock & Stoye, 2000; International Human Genome Sequencing Consortium, 2001). Many but not all HERVs have defects partly, or in all genes. However, despite millions of years since integration into the genome of the human ancestor, some HERV genes still have open reading frames (ORFs) and thus a possibility for protein expression. Expression of HERVs have been implicated for several positive physiological functions (Samuelson et al., 1990; Ting et al., 1992; Mi et al., 2000; Stoye & Coffin, 2000) and disease (Schulte et al., 1996; Perron et al., 1997; Christensen et al., 1998; Karlsson et al., 2001) yet not claiming the provirus as the etiological agent, but merely as a possible contributing factor and a marker of the diseases.

**Retroelements**

Retroelements are transposable genetic DNA sequences that pass an intermediate RNA stage in their replication cycle. These transposable retroelements (TEs) have been found in a wide range of organisms from prokaryotes to plants and animals (Mager & Henthorn, 1984; Garfinkel, 1994). The transposition involves synthesis of RNA that serves as a template for the RNA dependent DNA polymerase, also referred to as the reverse transcriptase (RT) (Baltimore, 1970; Temin & Mizutani, 1970). The RNA is reverse transcribed into complementary DNA (cDNA) that may be integrated or re-integrated into the host genome. Such integrated DNA must not necessarily have a specific contribution to the host phenotype. As long as the amount of these “hitch-hiking” elements is tolerated in the host genome, they will continue to spread like “selfish genes” and can be compared to not too harmful parasites (Orgel & Crick, 1980).

The retroelements can be divided into two fractions based on their genomic structures and occurrence or absence of long terminal repeats (LTRs) (Figure 1). The retroposons lack LTRs and are represented by long and short interspersed nucleotide elements (LINEs and SINEs). The LTR containing group of TEs can further be divided into the retroviruses, which may be infectious, and the non-infectious retrotansposons. The major difference between these two LTR containing subgroups is the env gene. It is absent in the retrotansposons. Thus, they can only reintegrate into the same genome from which they were transcribed (Prak & Kazazian, 2000; Kazazian, 2004). At least 50% of the human genome consists of repeat sequences derived from TEs. These transposon derived repeats are also often referred to as interspersed repeats (Table 1) (International Human Genome Sequencing Consortium, 2001).
Figure 1. Schematic overview of retrotransposing elements. Bold arrows indicate the directions of the elements. Small arrows indicate LTR directions. Encircled P, indicate internal promoters. Abbreviations are as follows: PBS, Primer binding site; ψ, approximate site for packaging signal; gag, group specific antigen (MA, matrix; CA, capsid; NC. Nucleocapsid); pro and PR, protease; pol, polymerase (RT, reverse transcriptase; RH, RNAse H; IN, integrase); env, envelope (SU, surface unit; TM, transmembrane protein) (See the text below for details).

Table 1. Characteristics and genomic fractions of human retroelements.

<table>
<thead>
<tr>
<th>Genomic Structure</th>
<th>Designation</th>
<th>Length</th>
<th>Copies</th>
<th>Human genome fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without RT, with internal promoter</td>
<td>SINE</td>
<td>100-300 bp</td>
<td>1500000</td>
<td>13 %</td>
</tr>
<tr>
<td>With RT and internal promoter</td>
<td>Retroposon (LINE)</td>
<td>6-9 kb</td>
<td>850000</td>
<td>21 %</td>
</tr>
<tr>
<td>With RT and LTR. Without env</td>
<td>Retrotransposon</td>
<td>1.5-3 kb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>With RT, LTR and env</td>
<td>ERV</td>
<td>6-11 kb</td>
<td>450000</td>
<td>7.8 %</td>
</tr>
<tr>
<td>LTR</td>
<td>Solitary LTR</td>
<td>300-1000 bp</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Combined numbers for Retrotransposons, ERVs, and solitary LTRs.
**Processed pseudogenes**

Processed pseudogenes are reverse transcribed cellular messenger RNAs (mRNAs) that later are reintegrated into the host genome. The expression of a pseudogene is under the influence of up- and downstream located host genomic promoters and enhancers (Figure 1). The processed pseudogenes are generated by the long interspersed nucleotide elements (See below) (Esnault et al., 2000). Recently produced pseudogenes, i.e. remnants of HERVs, are the results of repeated and independent LINE-mediated retrotransposition of retroviral mRNA (Pavlicek et al., 2002). When a pseudogene occasionally has acquired a promoter and becomes actively transcribed, it is called a retrogene (Figure 1).

**Retroposons**

Short interspersed nucleotide elements (SINEs) are retroposons that may be included into the retrogenes as they carry an internal promoter, but lack the coding abilities (Figure 1). SINEs are about 100-300 bp long and constitute about 13 % of the human genome (Table 1) (International Human Genome Sequencing Consortium, 2001). In total, there are more than 1.5 million SINE copies in the human genome (International Human Genome Sequencing Consortium, 2001). One of the most studied SINEs is the Alu element, present in more than a million copies (Smit, 1996; International Human Genome Sequencing Consortium, 2001). Since Alus lack a functional RT, they are non-autonomous and thus not able to transpose by themselves.

Long interspersed nucleotide elements (LINEs, or usually L1) contain, besides an ORF, also an additional pol gene with similarities to RT (Figure 1). The LINEs can promote their own transposition and are the most active autonomous transposable elements in the human genome (Kazazian & Moran, 1998; Prak & Kazazian, 2000). There are 3 types of LINEs (named LINE-1, 2 and 3, which are 6-9 kb long) that represent around 850000 copies, or 21 %, of the human genome (Table 1) (International Human Genome Sequencing Consortium, 2001). The LINE-1 (L1) is the only actively transposable human element and have been shown to generate processed pseudogenes (See above) (Esnault et al., 2000). Since the LINEs share similarity with the SINEs in the region just upstream of the poly-adenylation signal, the RT encoded by the LINEs may mobilize the non-autonomous SINEs (Alus) in trans (Kazazian, 2004). Activity of the human L1 was demonstrated in rat neuronal progenitor cells. An effect in neuronal development and diversity in humans was postulated (Muotri et al., 2005). However, several questions, and if the process could be evolutionarily maintained, remain to be answered (Ostertag & Kazazian, 2005).
Retrotransposons

Retrotransposons hold their promoter elements within their LTRs. Besides an RT, similar to a pol gene, the retrotransposons also code for additional structural Gag (NC) like proteins. They can in theory form aggregates and particles. The major difference between retrotransposons and retroviruses is the env gene (Figure 1). Thus, retrotransposons are not able to infect new cells without a retroviral Env, complemented in trans. The retrotransposons comprise the two rather typical families, named gypsy/Ty3 and copia/Ty1 (Figure 1). They are similar in their gene setups, but differ in the internal pol gene arrangement. The copia/Ty1 IN is rearranged upstream of the RT, for a review see Boke and Stoye (1997). Further, the gypsy/Ty3 elements encode an env in ORF3. So far, no active LTR-retrotransposons have been found in the human genome (Prak & Kazazian, 2000). However, gypsy/Ty3 elements have been identified in several vertebrate classes, e.g. lampreys, bony fish, amphibians and reptiles, but were in the same study not identified in mammalian and avian taxa (Miller et al., 1999).

Retroviruses

Retroviruses are similar to retrotransposons, but have an additional env gene (Figure 1). They are approximately 6-11 kb long and constitute, together with retrotransposons and solitary LTRs, about 7-8% of the human genome (Table 1) (Bock & Stoye, 2000; International Human Genome Sequencing Consortium, 2001). Rodent Intracisternal A Particles (IAP) and Virus like 30s RNA elements (VL30) resemble the retrotransposons by their lack of a functional Env, but are still considered as defective retrovirus elements (Boeke & Stoye, 1997). Retroviruses, as infectious and defective proviruses, have been reported from most organisms in the form of HERVs (Martin et al., 1981; Mager & Henthorn, 1984; Urnovitz & Murphy, 1996; Tristem, 2000) to various mammalian, piscine, avian, reptilian, amphibian ERVs etc. (Tristem et al., 1996a; Tristem et al., 1996b; Martin et al., 1997; Herniou et al., 1998; Huder et al., 2002; Kambol et al., 2003; Shen & Steiner, 2004). The exogenous retroviruses (XRVs) can be divided into four morphological types and, together with the endogenous retroviruses (ERVs, see below), into seven retroviral genera based on sequence similarities (Table 2). ERVs can further also be grouped into three more loosely defined classes (ERV class I, II, and III). The retroviral classification and taxonomy is discussed separately below.
Table 2. RNA reverse transcribing viruses.

<table>
<thead>
<tr>
<th>Genus</th>
<th>Type</th>
<th>Representing Virus</th>
<th>Organisation</th>
<th>Major Human groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>Avian type-C</td>
<td>ALV Avian Leukemia virus</td>
<td>Simple</td>
<td></td>
</tr>
<tr>
<td>Beta</td>
<td>Mammalian type-B, D</td>
<td>MMTV Mouse Mammary Tumour Virus</td>
<td>Simple/Complex</td>
<td>HERV-K(HML2)</td>
</tr>
<tr>
<td>Gamma</td>
<td>Mammalian type-C</td>
<td>MLV Mouse Leukemia Virus</td>
<td>Simple</td>
<td>HERV-E, HERV-F, HERV-I, HERV-T, HERV-W</td>
</tr>
<tr>
<td>Delta</td>
<td>Type-C like2</td>
<td>BLV Bovine Leukemia Virus</td>
<td>Complex</td>
<td>HTLV-1, HTLV-2</td>
</tr>
<tr>
<td>Epsilon</td>
<td>Type-C</td>
<td>WDSV Walleye Dermal Sarcoma Virus</td>
<td>Complex</td>
<td></td>
</tr>
<tr>
<td>Spuma</td>
<td>Type-C like3</td>
<td>HSRV Human Spuma Retrovirus</td>
<td>Complex</td>
<td>HSRV, HERV-L</td>
</tr>
</tbody>
</table>

1. HERV-K(HML2) is complex due to accessory genes rec or np9.
2. HTLV/BLV morphology similar to that of type-C.
3. Spuma morphology is similar to that of type-C, but with prominent surface spikes.

Retroviral genome structure

A typical replication competent provirus is about 6-11 kb in size and consists mainly of the coding regions gag, pro, pol and env, flanked on both 5′- and 3′-ends by LTRs (Figure 1). The retroviruses are either simple or complex, depending on their genomic structures (Table 2). Simple retroviruses are only maintaining the most essential genes, gag, pro, pol and env, coding for the virion proteins. The complex retroviruses have evolved to use accessory genes and their gene products. The accessory genes are overlapping with the essential genes, although usually in different reading frames, thus keeping the retroviral genome compact (Petropoulos, 1997; Vogt, 1997a; Vogt, 1997b).

The additional proteins can act in different cellular regulation processes and enhance the retrovirus replication efficiency. The complex HIV employs a system that includes the accessory protein Rev. The Rev binds full-length RNA transcripts or partially spliced subgenomic RNA (e.g. env), in trans to its RRE (Rev-response element) and export it from the nucleus to the cytoplasm. Thus, HIV avoids degradation of the unspliced RNA. The complex HTLV uses a similar trans-acting system involving Rex and Tax. For a review, see Rabson and Graves (1997).

The integrated provirus contains two LTRs, which have the characteristic start (TG…) and stop (…CA) sequences (Shimotohno et al., 1980). The LTR is composed of the unique U3 region, derived from the 3′end of the viral RNA, and U5 element from the 5′ end of the same RNA. The unique sequences (U3 and U5) are separated by a repeat segment (R), which is derived from a sequence present at both 5′ and 3′ end of the viral RNA. The U3 may vary in length and contains binding sites for different cellular transcription factors for enhancing and promoting proviral transcription. Several studies have been conducted on the LTRs of HERVs describing transcription factor (TF) binding sites and other important sequences like the TATA-box coupled with a GC/GT-box for transcriptional start and the AATAAA signalling poly-adenylation important for termination and transportation (Mager, 1989; Nelson et al., 1996; Sjottem et al., 1996; Anderssen et al., 1997).
1997; Sjøttem et al., 1997; Schon et al., 2001). The TATA-box is located in the U3 and the AATAAAA is located in the R region (Rabson & Graves, 1997). The viral transcription usually starts at the 5' U3-R junction, where the CAP-site is located. At the time of viral integration into the host genome, the LTRs are identical and may thus be used in estimation of the proviral age. Analysis of acquired proviral mutations can be approximated to a time scale.

The region following the 5’LTR and ranging to the start for the group specific antigen (gag), is called leader region (L) or 5’-untranslated region (5'UTR). This region contains the primer binding site (PBS), which is localized just downstream of the 5’LTR. The PBS is 18 nucleotides long and complementary to the 3’-sequence of the host transfer RNA (tRNA), which is used as a primer in the reverse transcription initiation. The PBS sequence has also been useful to name and group ERVs. Thus, members of the HERV-H have a PBS complementary to the histidine-tRNA, and most members of the HERV-K(HML) group have a Lysine-tRNA PBS. The L-region also contains the packaging signal (ψ) and “kissing loops” for dimerization of the two RNA copies that are packed into the virion. The major splice donor site (SD) used for subgenomic splicing of env, and also accessory proteins of complex retroviruses, is usually situated just downstream of the PBS. The L-region of murine leukemia virus (MLV) has been shown to enclose an alternative gag start (CUG) just upstream of the normal gag start (AUG), thus producing a larger glycosylated “glyco-Gag” (Fan et al., 1983). An internal ribosome entry site (IRES) used for alternative translation of initiation has also been reported for MLV within the L-region (Berlioz & Darlix, 1995; Vagner et al., 1995).

The gag is, after pol, the most conserved gene in the retroviral genome and codes for a polyprotein that is cleaved into the three structural proteins forming the inner part of the virion: MA (matrix), CA (capsid) and NC (nucleocapsid). The genetic arrangement is a necessity for the translated proteins to interact in a specified order, and to guide the next proteins into position in order to assemble the virion correctly from the outside to the inside (Coffin, 1994). The MA attaches to the cell membrane and most retroviral MA need modifications by addition of a myristic acid group to their N-terminus (Henderson et al., 1983; Rein et al., 1986).

In the MA-CA junction, there is in several retroviruses often a proline rich region including a PPPY motif, that encodes a so-called “late” function (Wills et al., 1994). This late assembly domain has been shown in both RSV (PPPY) and the corresponding late domain in HIV-1 NC (P(T/S)AP), to be involved in the recruitment of ubiquitin ligase (Patnaik et al., 2000; Strack et al., 2000; Vogt, 2000). One of the cellular roles of poly-ubiquitination of
proteins is to mark them for proteolysis. This may also be used by the virus in the budding process. Further, the late assembly domain has been shown to interact with the Nedd4 family of ubiquitin ligases that is required for the budding process in RSV p2b (Kikonyogo et al., 2001). The N-terminal CA, shaping the virion core structure, has a 20 amino acid major homology region (MHR. See also figure 9, paper III below), which is similarly conserved between most retroviruses (Benit et al., 1997; Petropoulos, 1997).

The C-terminus of NC contains a highly conserved region with one or more zinc fingers (CCHC, consensus: CX₂CXX₄HX₄C). The CCHC motif, conserved among most retroviruses but absent in spumaviruses, coordinates Zn²⁺ ions (Chance et al., 1992). It has a role in binding other proteins to nucleic acids. It probably interacts with the RNA, targeting the packaging signal during the capsidation.

The protease gene (pro), located between gag and the polymerase (pol), encodes a protein (PR) responsible for cleaving the different retroviral polyproteins into active subunits. The HIV PR has been subjected to comprehensive studies, including crystallography, in attempts to find alternative strategies to stop HIV infection and thus prevent the progression of AIDS. The pol is encoding a polyprotein, cleaved into the active subunits RT (reverse transcriptase), RH (RNase H) and IN (integrase). It is the most conserved gene among retroviruses. The RT is more error prone than DNA polymerases, which may be beneficial for the retrovirus in creating genetic variation. Another tendency for RT is to jump between the two retroviral genomic RNA strands during reverse transcription (Temin, 1993). The implications for this would be an increased variation, but also to increase the probability for the retrovirus to overcome eventual defects in one of the RNA strands (See “Break out” below). In MLV, with gag and pro/pol in the same reading frame, the Gag-Pol fusion protein was suggested to be generated by a readthrough of the UAG stop codon after gag, probably by misincorporation of an amino acid (Philipson et al., 1978; Coffin, 1985). This was later shown to also involve pseudoknots that most probably stall the ribosome and facilitate the readthrough from gag to pro/pol during translation (Wills et al., 1991).

The envelope gene (env) is transcribed through subgenomic splicing from the major SD (See above), and is coding for the Env polyprotein. The Env is subjected to proteolytic cleavage into the surface unit (SU) and transmembrane protein (TM). Env proteins are usually referred to by their molecular weight, e.g. gp120 for HIV SU. The SU is, unlike TM, usually glycosylated and hence the “gp” prefix. Since Env is used for fusion and retroviral entry, it also determines which cell type the retrovirus infects. There are two known general classes of viral membrane fusion proteins, class I and II
Jardetzky & Lamb, 2004), not to mistake for the ERV classes. The class II fusion proteins do not occur in retroviruses. The class I fusion proteins, including the retroviral Envs, form SU and TM homotrimers. Briefly, in a relaxed state the homotrimer binds its cell surface receptor. The binding, or exposure to low pH, triggers an elongated formation of the Env-receptor complex. When several trimers bind in vicinity to each other, protein refolding occurs to release energy. The viral and cell membranes are brought into proximity and the fusion takes place. The retroviral Env is similar to the best studied viral Env glycoprotein: Influenza hemagglutinin (HA), which is the model for the class I membrane fusion proteins (Jardetzky & Lamb, 2004). Analogous to the HA, the avian leukosis virus (ALV) Env has been shown to depend on low pH to undergo conformational changes leading to fusion (Mothes et al., 2000).

Other important Env properties, besides that of membrane fusion, is the demonstrated immunosuppressive domain (ISU) of the TM, also referred to as CKS-17 (Figure 2) (Cianciolo et al., 1985; Mangeney & Heidmann, 1998; Lindeskog et al., 1999; de Parseval et al., 2001; Mangeney et al., 2001).

![Figure 2. Schematic view of a gammaretrovirus with its immunosuppressive unit (ISU, containing the 17 amino acids conserved CKS17 motif) highlighted in TM (Adapted from Blomberg et al., 2005).](image)

The PPT (poly purine tract), that escapes the RNase H degradation during reverse transcription, immediately precedes the 3’LTR. The PPT can thus function as a primer for the synthesis of the second DNA strand (plus-strand strong-stop DNA), see Telesnitsky & Goff (1997) for a review.

**Retroviral replication**

The retrovirus life cycle differs from that of other organisms as they undergo the transformation in genetic material from RNA to DNA, and finally close
the replication cycle as RNA (Figure 3). The retroviral SU determines which cell to infect by binding to its specific receptor and sometimes co-receptors. The binding induces conformational changes in the SU and TM and forces TM into vicinity of the cell membrane, thus enabling fusion (Jardetzky & Lamb, 2004).

The virion core is released into the cytoplasm, where it is uncoated. In the cytoplasm, the retroviral RNA genome undergoes reverse transcription into dsDNA. The reverse transcription takes place in a protein and nucleic acids preintegration complex including the retroviral RT (Telesnitsky & Goff, 1997). The two identical flanking retroviral LTRs (U5-R-U3) are formed during the reverse transcription from being R-U5 in the retroviral RNA 5’ end and U3-R in the retroviral RNA 3’ end. The two identical RNA R sequences are necessary for the reverse transcription process (Telesnitsky & Goff, 1997). The identical LTRs at the integration moment can help us to estimate the age of endogenous proviruses, by examining the amount of mutations accumulated over time (See present investigation below). The close proximity between the 5’ ends of the two RNA copies, provides a chance for the RT to jump between the strands. Switching between the two strands may be an alternative for the retrovirus to eliminate deleterious mutations, and together with the error prone RT create more genomic variation (See nucleotide bias below and Katz & Skalka, 1990; Overbaugh & Bangham, 2001). Ultimately the jumping RT may lead to retroviral rescue and gradual self optimization (See “Break out” below and Blomberg et al., 2005).

The dsDNA is transported into the nucleus where it integrates into the chromosomal DNA. However, an obstacle for the retrovirus is the nuclear membrane that presents a blocking barrier for the preintegration complex to reach the chromosomal DNA targets. Simple retroviruses (e.g. MLV), without accessory genes and proteins that allow them to control the transport are depending on the cell cycle. The integration of a simple retrovirus has to take place during mitosis, when the cell nucleus is dissolved. Complex retroviruses (e.g. HIV) have, however, evolved strategies to control the nucleic acids transport and can thus integrate in non-dividing cells. The functions of the preintegration complexes are not yet entirely described and much work remains.

The integrated retrovirus is called a “provirus” and may be transcribed to form progeny RNA that is exported to the cytoplasm. Some RNA undergoes splicing and is transported to the cytoplasm where it is translated into structural and enzymatic proteins. Complex retroviruses like HIV and HTLV have evolved some of their accessory genes with similar functions (rev and rex, respectively) as integral parts of their transcription, promoting active transport of unspliced RNA from the nucleus (Rabson & Graves, 1997).
HERVs were long considered to be simple retroviruses. However, the betaretroviral HERV-K(HML2) was shown to have the accessory genes, rec and/or np9 (Lower et al., 1996; Armbruester et al., 2002; Armbruester et al., 2004), and is thus a complex retrovirus. Rec is at least functionally related to HIV Rev and HTLV Rex (Magin et al., 1999).

The structural proteins Gag and Env localize to the cell membrane. The two progeny RNA copies are incorporated into the virion together with the enzymatic Pro-Pol polyproteins. After budding from the cell membrane, the retrovirus matures as the polyproteins are cleaved into functional subunits.

Figure 3. The retroviral life cycle. (See the text for details).
Endogenous retroviruses (ERVs)

Endogenization

Endogenous retroviruses (ERVs) are genetic elements that reside as proviruses in their host’s genome. Retroviruses usually infect somatic cells and cannot transmit their proviral DNA to the host’s progeny. Occasionally a retrovirus infects a germ cell. If the germ cell survives, there is a chance that the acquired provirus will be passed to the offspring. The provirus is then inherited as a normal Mendelian gene (Stoye & Coffin, 1985). After the first human endogenous retroviruses were discovered (Martin et al., 1981; Bonner et al., 1982), the sequencing of the human genome (International Human Genome Sequencing Consortium, 2001) has shown that it is littered by these elements (Lower et al., 1996; Sverdlov, 2000; Tristem, 2000). Around 7-8% of the human genome has been estimated to be of retroviral origin (Bock & Stoye, 2000; International Human Genome Sequencing Consortium, 2001). Although HERVs with ORFs have been described, no actively autonomous retrotransposing HERVs have yet been observed (Mayer et al., 1999). However, the integrated provirus may contribute to the genomic plasticity (Lower et al., 1996), and as long as the cell can afford to maintain these integrated elements, these “selfish genes” will continue to exist and multiply (Orgel & Crick, 1980). A computer based analysis of the human genome sequence has shown a discrepancy in where the retroviral elements are integrated. All classes of LTR elements and LINES (Figure 1) were under-represented within and in the vicinity of genes, probably because of their potential to influence transcription (Medstrand et al., 2002). This phenomenon was also demonstrated in an in vivo assay, where proteins that bound in the vicinity to genes suppressed the retroviral integration (Weidhaas et al., 2000).

Biological relevance of ERVs

Promoters and enhancers

The retroviral elements that integrate in the vicinity of genes may interact with their host cell genome in various ways. The numerous potential binding sites for transcription factors in the LTRs may influence the transcriptional activity of nearby genes. An example is the expression of amylase in the human parotid glands, where integration of HERV-E in reversed orientation upstream of the amylase gene promotes expression (Samuelson et al., 1990; Ting et al., 1992). This bidirectional promoter activity has also been observed in the large HERV-H group, that had strong promoter activities in several cell lines (Feuchter & Mager, 1990). The high number of HERV-H
integrations in the human genome (926 and additional solitary LTRs, paper II), increases the likelihood for effects on adjacent genes.

In human malignant trophoblasts (e.g. choriocarcinoma), HERV-E integration into the growth factor gene pleiotropin (PTN) has generated cell type specific promoter activity (Schulte et al., 1996; Schulte et al., 2000). Cell type specific expressions in a variety of human cells under the influence of LTR promoters have also been described for several other HERVs (Schon et al., 2001). LTR promoters may further enhance the transcription from a native promoter. Presence of a HERV-E LTR increases the native promoter activity and expression of Apolipoprotein C-I (Medstrand et al., 2001).

Repetitive elements have been found to integrate in a vast amount of human 5’UTRs (untranslated regions). A study showed that 4 % of the human 5’UTRs harbour Alus (SINE, figure 1) and that many genes may be under influence of these integrations. The same study demonstrated a tendency for these Alus to increase transcription efficiency while translation efficiency was decreased (Landry et al., 2001). In a following study, TEs were shown to promote regulatory functions. They aggregate in vicinity to genes with low conservation, such as the mammalian genes involved in immunity. In contrast, TEs had a low tendency to aggregate next to fundamental genes with high conservation e.g. metabolism and cellular structure genes. The conclusion was that TEs have had a major impact on the evolution of gene families in mammals (van de Lagemaat et al., 2003).

The human β1,3-galactosyltransferase 5, thought to be responsible for type-1 Lewis antigen that may be involved in tumour metastasis in some colorectal and pancreatic cells, have been found to use a HERV-L LTR in its first exon (Dunn et al., 2003). For a more extensive review of the TEs influences on the human genome, see Brosius (1999).

**Alternative and intergenic splicing**

Integration of retroviral elements, and especially the LTRs, may act as switches for alternative splicing. The leptin obesity hormone receptor (OBR) has been found in two variants that differ in size. The cause was mapped to alternative splicing mediated by the HERV-K LTR, thus generating a shorter gene transcript (Kapitonov & Jurka, 1999). Intergenic splicing can also occur through initiation by an LTR upstream of the gene. The HERV-H has demonstrated such properties and its LTR initiates a phospholipase A2 related gene (used for digestion in pancreatic juices), generating the single copy gene PLA2L (PLA2 like) which is expressed in human teratocarcinoma cells (Feuchter-Murthy et al., 1993). It was shown that the last two-thirds of PLA2L were derived from the human orthologue of mouse Otoconin-90 (PLA2L/OC90), a major protein in the otoconia of the inner ear, which is
vital for the sense of gravity (Wang et al., 1998). The transcript was found to be the product of intergenic splicing between a HERV-H element and the two downstream genes normally independently expressed from different promoters. The transcript starts in the HERV-H LTR and uses the major SD to splice into HHLA1 (HERV-H LTR Associated gene) downstream of the provirus, followed by a second splicing into the PLA2L/OC90 (Kowalski et al., 1999) (See also figure 8 in the present investigation).

The normal function for alternative splicing is to provide a possibility for complex retroviruses to maintain accessory genes (e.g. HIV Rev and HTLV Rex, see above and Rabson & Graves, 1997), encoded in different reading frames and still keep the genome size small enough for packaging into the virions. HERVs also have accessory proteins. HERV-K(HML2.HOM) has an additional gene (cORF) encoding Rec (Figure 4), which shares similarity with the HIV Rev/Rex function in transport of unspliced retroviral RNA from the nucleus to the cytoplasm (See retroviral replication above and Magin et al., 1999). An analysis of the human genome revealed that 11/17 HERV-K had intact cORFs but it could only be produced from 7 loci (Mayer et al., 2004). Further, a deletion in the cORF produces a 9 kDa protein, (Np9, figure 4) that locates in the nucleus and is expressed in various tumour tissues (Armbruester et al., 2002).

Figure 4. Spliced mRNA transcripts in types of HERV-K(HML2). In addition to full-length mRNA and spliced env mRNA, type 2 produces rec (cORF) after splicing and frameshift from ORF-1 to ORF-2 (numbers over mRNA). Type 1 has a 292-bp deletion and produces np9 instead after splicing and frameshifting from ORF-1 to ORF-3. (Adapted from Blomberg et al., 2005).
Impact of retroviral Env

Viral gene products are significant for the human cell. The human protein “Syncytin”, which is identical to the well described HERV-W Env, may be an example of a viral gene that has been adapted to serve important physiological functions in its host. Syncytin is found in the trophoblasts of the placenta, where it mediates cell fusion and syncytia formation. This syncytiotrophoblast layer is important to protect the foetus from maternal immune response during development (Blond et al., 2000; Mi et al., 2000; An et al., 2001; Frendo et al., 2003). Recently the human genome was screened for other fusogenic candidates. After cloning and expression of 16 candidates, it was found that also HERV-FRD had the same properties as shown for Syncytin. It was subsequently called “Syncytin 2” (Blaise et al., 2003). Shortly after, it was proved that the HERV-W provirus with conserved ORF at locus ERVWE1, encodes the Syncytin 1 (EnvW, derived from HERV-W). It was proposed that it has been subjected to positive selection during the primate evolution for its role in placentation (Mallet et al., 2004). Further, the Syncytin-1 was functional in a pseudotyped MLV vector assay (An et al., 2001), and could be cleaved into a TM subunit and a glycosylated SU. The SU and TM subunits were found in homotrimerers (Cheynet et al., 2005), as expected for a class I fusion protein (Jardetzky & Lamb, 2004).

Recently two new fusogenic Envs, unrelated to the human Syncytin-1 (EnvW, derived from HERV-W) and Syncytin-2 (EnvFRD, derived from HERV-FRD) were described in the muridae (Dupressoir et al., 2005). These two murine Syncytin-A and B, indicate a positive selection for a function similar to that of the human Syncytins. Thus, presenting at least two cases where ERV genes may have been acquired independently for a common physiological role (Dupressoir et al., 2005). Since these findings, two additional human genes (envV and envP(b)) with partial homology to the envW and envFRD have been described. Interestingly, both EnvV and EnvP(b) were phylogenetically separate from the EnvW and EnvFRD. EnvV was expressed in placenta but had no fusogenic properties. The EnvP(b), however, had fusogenic properties, but did not show any upregulated expression in any of the tissues tested (Blaise et al., 2005). Thus, fusogenic ERV derived proteins are not limited to placental expression, and further experiments should be conducted to elucidate the possible physiological effects.

The HERV-H has, despite numerous damaging mutations, env ORFs, in a total of three loci, on chromosome 1 and 2 (Paper I and de Parseval et al., 2001) of 91 env containing loci (Paper II). The function of HERV-H Env is unknown, however, Env has been observed to have immunosuppressive properties (See ISU/CKS17 figure 2) (de Parseval et al., 2001; Mangeney et al., 2001). It has been associated with cancer (See below).
Another ERV with \textit{env} ORF is the ERV3 (also referred to as HERV-R) at chromosome 7q11 (Andersson et al., 2005). ERV3 has been postulated a contribution to cellular differentiation. \textit{In situ} hybridization (ISH) experiments, using ERV3 \textit{env} specific probes, showed transcriptional activity in several foetal tissues with high levels in the developing adrenal gland, compared to other tissues (Andersson et al., 2002). Upregulation of ERV3 RNA and protein expression has also later been observed in adrenal glands and placenta (Andersson et al., 2005). Even brain (whole) tissue had upregulated ERV3 RNA expression, detected by quantitative real-time PCR targeting the \textit{env} SU. Using ISH, Andersson et al. also found expression in an array of other tissues including prostate, testis and sebaceous glands of the skin, thus confirming previous reports. An interesting issue is a polymorphism in 1% of the Caucasian population, resulting in a premature ERV3 \textit{env} stop codon (de Parseval & Heidmann, 1998). Unless a truncated protein is sufficient, it is therefore hard to conceive a primary function for ERV3 Env during foetal development.

Envelope glycoproteins mediate receptor interference as a strategy to prevent superinfection, i.e. two or more retroviruses infecting the same cell (Boeke & Stoye, 1997). However, the cell is still susceptible for infection of viruses that use different receptors. Eleven retroviral receptor groups have been identified in human cells, Receptor interference have been suggested to occur predominantly in cell cultures where the chronically infected cells express sufficient amounts of Env to block their cognate receptors (Sommerfelt, 1999). This was demonstrated, \textit{in vitro}, with transfected HERV-W that protects cells against Spleen necrosis virus (SNV) infection (Ponferrada et al., 2003).

\textbf{Retroviral restriction}

In MLV, a single amino acid residue in CA determines the virus tropism for BALB/c (B-tropic) or NIH-3T3 (N-tropic) cells and integration restriction, postentry but preintegration, by the cytoplasmic \textit{Fv1}. The \textit{fv1} gene was demonstrated by cloning and sequencing to have derived from the \textit{gag} of an endogenous MLV (Best et al., 1996; Benit et al., 1997). A wide array of mammalian cell lines, including human, had \textit{Fv1} retrovirus like restriction functions, when infected with an N-tropic vector pseudotyped with the broadly binding vesicular stomatitis virus G (VSV-G) envelope. The specificity of \textit{Fv1} restriction mapped to the CA (position 110) and the experiments indicated a block in the synthesis step or in the accumulation of viral DNA (Towers et al., 2000).

Recently, a novel innate retroviral defence mechanism was identified where the cytoplasmic TRIM5\(\alpha\) of old world rhesus monkeys restricted HIV-1 infection (Stremlau et al., 2004). Simian immunodeficiency virus (SIV) was
less susceptible to the rhesus TRIM5α-mediated block. The human TRIM5α does the opposite. Further, the TRIM5α effect was saturable as previously indicated (Stoye, 2002). The similarity to the Fv1 restriction was concluded by demonstrating that TRIM5α (from human, rhesus monkey and African green monkey) also restricted N-tropic MLV (Yap et al., 2004). The retroviral restriction has been mapped to a segment consisting of a 13 amino acids patch (in the C-terminal part, SPRY domain) that has been under positive selection that predate the origin of lentiviruses (Sawyer et al., 2005).

Retroviral restriction in mice can also occur at the entry, where Fv4 blocks the cell surface receptor from interaction and subsequent fusion with the ecotropic (i.e. only infecting mice) MLV (Rosenberg & Jolicoeur, 1997). The Fv4 restriction (blocking) mechanism was shown to depend on a defective endogenized Fv4-Env compared to the functional ecotropic MLV. A single non-synonymous substitution for arginine in Fv4 Env (R491) compared to a glycine (G491) in the ecotropic MLV Env, was sufficient to inhibit the TM interaction and fusion with the receptor (Taylor et al., 2001).

Receptor Blocking is also utilized by sheep cells, where endogenous JSRV (enJSRV) restricts the entry of exogenous (infectious) JSRV (Palmarini et al., 2004). A second restriction in sheep occurs in the cytoplasm. The enJSRV produces a Gag that does not permit the virus to exit the cell. When enJSRV is expressed in the same cell as exogenous JSRV, hybrid Gags are formed, thus preventing both viruses from exit of the cell (Palmarini et al., 2004). Similarly to the single non-synonymous substitution for Fv4 (above), the enJSRV Gag restriction depends on a single R21W mutation compared to JSRV (Mura et al., 2004). The enJSRV is, in addition to the human syncytin, another example of an ERV with ORF that is probably maintained by the cell because of its beneficial roles.

An additional innate cell restriction mechanism against lentiviral HIV is that of APOBEC3G. The APOBEC3G promotes G to A hypermutations in the HIV by deamination of cytidine in the negative strand synthesis mutating to a thymidine (Mangeat et al., 2003; Vartanian et al., 2003). This turned out to be the probable cause of the pronounced nucleotide bias in HIV (Table in appendix A) observed earlier (Vartanian et al., 1994; Berkhout et al., 2002). The APOBEC3G also restricted ERV retrotransposition of Env deficient mouse MusD and IAP elements (Esnault et al., 2005).

**Cancer**

Historically, cancer has been the dominant disease associated with retroviruses. The first cancer associated viruses were avian and were discovered in the beginning of the 20th century. Oncogenes carried by viruses (v-onc) were discovered and soon after, also cellular proto-oncogenes (c-onc) were
found. For a review, see Rosenberg & Jolicoeur (1997). Despite that Dr. Bittner, in 1936, had detected the transmission of mammary carcinoma in mice, it was still believed that leukemia in mice, and other leukemias were native genetic disorders. In 1951, Dr. Gross made the discovery that the disorder was also infectious. Proviruses that caused mouse leukemia could be passed naturally from one generation to the next. He used embryo cell suspensions from a mouse of a leukemia inbred strain (Ak mouse) and inoculated a leukemia free strain. Thus, he proved that "spontaneous" leukemia may be caused by an agent that is transmitted from one generation to the next (Gross, 1951). At present, the focus on HERVs and cancer is increasing. The HERV-K(HML2.HOM), which is the most intact human provirus with full ORFs for all genes, was found to express Gag proteins. Antibodies against Gag could be measured in high titers in seminoma patients at the time of tumour detection. It was further established that Gag was expressed in the cytoplasm of the tumour cells (Sauter et al., 1995). The antibody response against Human teratocarcinoma derived virus (HTDV, i.e. HML2) Gag and Env precursors could be characterized using immunofluorescence and electron microscopy in teratocarcinoma cell lines. However, no processed Gag protein could be detected (Boller et al., 1997). Spliced env transcripts from the same virus were detected in human breast cancer, but not in healthy controls (Wang-Johanning et al., 2003b). Supernatants from human melanoma cells had RT activity and retrovirus like particles were detected by electron microscopy. Particles containing mature Gag and Env proteins could be isolated (Muster et al., 2003). The two HERV-K108 (Barbulescu et al., 1999) and K113 (Turner et al., 2001) were identified. Thus, without establishing pathogenicity, melanoma-associated retroviruses in the recently active HERV-K(HML2) group (See paper V) are present as quasispecies. Added to that, is the interesting polymorphism for HERV-K113, which thus has not been fixed in the population (Turner et al., 2001) (See below). These findings open for the discussion of master elements (See below, paper II and Blomberg et al., 2005).

In another study, the HERV-E env was expressed in 19/49 prostate carcinoma tissues compared to 1/18 in normal controls (Wang-Johanning et al., 2003a). HERV-E expression was also significant in the choriocarcinoma growth, influencing the PTN (See promoters and enhancers above) (Schulte et al., 1996; Schulte et al., 2000).

Also HERV-H is connected to cancer. Expression of the HERV-H Env, was determined in lung squamous carcinoma (Hirose et al., 1993). The immunosuppressive properties of the TM (Figure 2), have now been proved, not only in vitro (de Parseval et al., 2001; Mangeney et al., 2001), but also in vivo as
the related MLV Env with a similar ISU as the HERV-H TM, was required for melanoma tumour growth in a mouse model (Mangeney et al., 2005).

Bioinformatic approaches, including in silico analysis of expression by scanning databases for reported expressed sequence tags (ESTs), have proved useful as complements to the laboratory experimental data. An extensive study matched HERV proviruses to ESTs and found that HERV-H was more frequently detected in ESTs from cancer tissues compared to ESTs derived from normal tissues. HERV-H had higher expression in stomach, colon, prostate, and testis tumour compared to other HERVs (See paper III and Stauffer et al., 2004). Expression of HERV-H gag transcripts have also been detected in human leukemia cell lines, and was demonstrated in a luciferase assay to not depend on the cell specific activity of transcription factors (Patzke et al., 2002).

To conclude, using a wide range of techniques, proviruses from the most distantly related genera (gamma and beta, figure 5) have been more or less firmly associated with cancer. This is not only a human phenomenon, but is also found in other mammals, e.g. lung tumours in sheep caused by the Jakobsiekte retrovirus (JSRV) (Wootton et al., 2005). It is interesting that various retroviral genomic portions are expressed. This creates a suspicion that upregulation is a secondary effect of the disease. In a murine model of cancer cachexia, several retroviral sequences were upregulated (Monitto et al., 2001). The ERVs may have been conserved against disrupting mutations during evolution because of beneficial effect to the cells.

Neurodegenerative disorders and autoimmunity
Numerous reports have connected HERVs to neurodegenerative diseases such as multiple sclerosis (MS) and schizophrenia. Multiple sclerosis associated retrovirus (MSRV) was found by RT-PCR (with degenerated broadly detecting “PAN primers”) to be expressed in cell lines and plasma from MS patients. Fractions from a sucrose density gradient had a peak for RT activity around the corresponding size for a retrovirus (1.16-1.18 g sucrose/ml) (Perron et al., 1997). The MSRV was characterized after cloning and sequencing, associating it to HERV-W (Blond et al., 1999; Komurian-Pradel et al., 1999). Recently, the HERV-W Env (Syncytin) expression was found to be upregulated in demyelinating brain tissues of MS patients (Antony et al., 2004). It was further speculated that proinflammatory properties add a second possible role for Syncytin, next to that discussed for placenta formation during pregnancy (Blond et al., 2000; Mi et al., 2000; Frendo et al., 2003).

In contrast to these findings, another group found HERV-H particles with RT activity in MS samples (Christensen et al., 1998). They also found that the HERV-H particle expression was tissue specific to MS samples com-
pared to other autoimmune diseases and healthy controls. The particles had an RT peak at 1.24 g/ml in their sucrose density gradient. Virus-like particles were detected using electron microscopy (Moller-Larsen & Christensen, 1998). Transmissibility was investigated using co-cultivation assays. The results indicated that HERV-H was transmissible at a low level (Christensen et al., 2002). However, HERV-H env RNA expression \textit{in vivo} was not elevated in MS patient samples (Antony et al., 2004). Thus, so far two retroviruses have been erratically associated with MS. An explanation could be that several proviral loci are activated. It is also possible that the change in HERV expression may be the result, and not the cause, of inflammatory disease within the brain and increased macrophage activity (Johnston et al., 2001).

Schizophrenia is, akin to MS, a neurodegenerative disorder, except for a central nervous system (CNS) inflammation found in MS. An increase in HERV-W RNA expression has been shown, using representational difference analysis (RDA), in monozygotic twin pairs discordant for schizophrenia patients compared to those of healthy controls (Debrinker et al., 1999). RNA containing HERV-W particles have been detected in cerebrospinal fluid of newly debuted schizophrenia patients. However, also other retrovirus transcripts were found, albeit in lower amounts (Karlsson et al., 2001). Although low expressions were detected, strength in these experiments was the use of Owl monkey kidney cells. This New world monkey does not have many of the more recent HERV integrations. The possible ERV involvement in schizophrenia and also MS is interesting since the HERV-W receptor is the transporter protein for glutamate, an important brain signal molecule (Lavillette et al., 2002). Thus, the connection between HERV-W and the neurodegenerative diseases MS and schizophrenia is theoretically realistic.

However, recently we studied an array of human tissues, using both broadly amplifying and specific \textit{pol}-based quantitative real-time PCR methods (Forsman et al., 2005). The measurements were compared to the housekeeping gene for Histone 3.3, which is evenly expressed in many human tissues (Andersson et al., 2005). We did not find any elevated HERV-W RNA expression in brain tissue, but we found HERV-H RNA expression to be significantly higher. Taking the different results together, we now have a good array of sensitive methods to overcome the bias in detection, and a control of the biased findings of claimed HERV-H (Christensen et al., 1998; Moller-Larsen & Christensen, 1998; Christensen et al., 2002) and HERV-W (Perron et al., 1997) expressions in MS samples should be conducted.

Systemic lupus erythematosus (SLE) is an autoimmune disease resulting in production of autoantibodies as an effect of loosing the ability to recognize
and build a tolerance to self antigens. Patients suffering from SLE have increased amounts of IgG reactive to different HERV Gag and Env derived peptides (Blomberg et al., 1994). Also HERV-E gag transcripts have been described in SLE (reviewed in Sekigawa et al., 2001). The retroviral contribution to SLE is thus not clear. However, using a mouse model of lupus nephritis, an ERV was the most highly expressed gene in the kidneys (Alexander et al., 2002), and supports the involvement of ERVs in SLE. Which effect(s) ERVs have in the SLE or if it is a secondary effect caused by the disease progression remains to be tested.

Retroviral classification and taxonomy

Morphology based classification

Early classification of retroviruses was based on the particle morphology observed in electron microscopy (The international Committee on Taxonomy of Viruses [http://www.ncbi.nlm.nih.gov/ICTV/]). The Retroviruses (exogenous) were classified into four types (Coffin, 1994).

- **A-types** are also known as intracisternal A particles, IAP. They are non-enveloped, immature particles and are only detected inside cells, believed to result from endogenous retrovirus-like genetic elements.

- **B-types** are represented by the Mouse mammary tumor virus (MMTV) (Table 2). They are enveloped and form extracellular particles with a round condensed, acentric core and prominent envelope spikes.

- **C-types** are represented by most mammalian, e.g. Murine Leukemia Virus (MLV) and avian retroviruses e.g. Avian Leukosis Virus (ALV) (Table 2). Their morphologies are similar to the B-types, but with a round sometimes angular central core and barely visible spikes. Spumaviruses also resemble the C-type but with a less condensed core and more prominent envelope spikes.

- **D-types** are represented by Mason-Pfizer monkey virus (MPMV). They are usually slightly larger (to 120 nm) and have a bar shaped core and less prominent envelope spikes.
The HIV-1 with its distinct morphology of a cone shaped core, has not been designated one of the four types above, although it resembles the C-type in budding (Petropoulos, 1997).

Sequence based classification (XRV and ERV)

The genome sequencing projects, human (International Human Genome Sequencing Consortium, 2001), chimpanzee (available publicly at [http://genome.ucsc.edu/]), and chicken (International Chicken Genome Sequencing Consortium, 2004), have resulted in that numerous previously not described retroviruses (proviruses) and related retrotransposing elements with clinical implications have been found (Urnovitz & Murphy, 1996; Kazazian, 1998; Larsson & Andersson, 1998; Lower, 1999). The principle of retroviral sequence comparison (Johnson & Coffin, 1999) has made large scale phylogeny and classification possible. Sequencing of retroviruses (XRVs and ERVs), have updated the modern classification to seven retrovirus genera: alpha, beta, gamma, delta, epsilon, lenti and spuma (Table 2).

The ERVs, with a broad sequence variation and subjected to numerous mutations during the evolution, can be difficult to assign to groups and ultimately a suitable taxonomy. Thus, ERVs can alternatively be divided into more loosely defined ERV classes I-III (Wilkinson et al., 1994; Lindeskog, 1999; International Human Genome Sequencing Consortium, 2001; Mager & Medstrand, 2003). To date 31 ERV families have been documented that are distributed into 23 class I, 4 class II and 4 class III elements (Katzourakis & Tristem, 2005).

Currently, the retroviral PBS, complementary to the host tRNA, serves to name and group ERVs into families (Urnovitz & Murphy, 1996; Tristem, 2000; Gifford & Tristem, 2003). However, this nomenclature needs revision, since a group of HERVs may share high similarity within their genes, while having different PBS-types (See paper II and Mager & Medstrand, 2003).

The LTRs have been used to study and group retroviral elements with limited success (Goodchild et al., 1993; Anderssen et al., 1997; Johnson & Coffin, 1999; Costas & Naveira, 2000; Costas, 2002; Hughes & Coffin, 2004). A limitation in using the nucleotide sequences of proviruses is the wide diversity. Nucleotide sequences are in general more difficult to compare than protein sequences. In nucleotide alignments there are only identities or non-identities of the four nucleotides (A, T, G and C) between the sequences. In protein alignments, there are more possibilities for sequence match since amino acids, besides identity, also can share structural similarities e.g. leu-
cine and iso-leucine. Thus, protein sequences are preferable when studying
distantly related sequences.

Env protein sequences have been used in phylogenetic reconstructions and
evolutionary studies (Benit et al., 2001; de Parseval et al., 2003). However,
generation of sequence diversity in HIV Env is an example of the retroviral
strategy to escape the immune system (Gaschen et al., 2002). Thus, phylog-
enies using Env are more useful in comparing closely related sequences than
diverse sequences. The TM is more conserved than the SU, and is therefore
better for alignments of more distantly related sequences (See Benit et al.,
2001). The ratio between synonymous (ds, nucleotide substitutions not caus-
ing change in amino acid, per site) and non-synonymous mutations (da [or
dn], nucleotide substitutions causing change in amino acid, per site) can be
useful to investigate a possible selective pressure on the ORF. Sequences
with maintained ORF and ds/da>1 are under positive selection and negative
selection if ds/da<1 (Li, 1997). Thus, sequences with important structural or
enzymatic functions would be under positive selection (ds/da>1). The Env
SU could on the other hand in some regions be under negative selection
(ds/da<1), as a strategy for the retrovirus to escape recognition by the im-
une system.

In several studies, pol derived sequences (nucleic acids and alternatively
protein sequences) have been used to reconstruct phylogenetic relationships
between different ERVs, XRVs and related elements (Papers II, IV and V)
(Xiong & Eickbush, 1988; Andersson et al., 1999; Tristem, 2000; Andersson
et al., 2005; Greenwood et al., 2005; Oja et al., 2005). The pol is the most
conserved retroviral gene, probably due to its fundamental properties and
role in the retroviral replication. Among seven conserved RT (RT1-7) motifs
(Xiong & Eickbush, 1988), the LPQG (in RT4) and YXDD (in RT5) (Shih et
al., 1989), where X is an arbitrary amino acid, are usually well suited for
phylogenetic inference. These conserved motifs were also used to develop
broadly targeting PCRs (Shih et al., 1989; Medstrand et al., 1992; Seifarth et
al., 1995; Forsman et al., 2005). However, these methods have a relatively
low sensitivity (Seifarth et al., 1995; Forsman et al., 2005). An explanation
for the low sensitivities may be deleterious mutations in RT, exemplified by
the RTV1H2-like subgroup of HERV-H (Paper II). This may also cause
trouble in Pol phylogenetic inferences. Nevertheless, expression analyses
using the broadly amplifying PCRs, and later sequencing, have added to the
accumulated retroviral sequence knowledge.

In order to track retroviral data in various genomes, we developed a bioin-
formatic tool RetroTector© (Paper VII, not included into this thesis). Briefly,
RetroTector© recognizes conserved consensus motifs and reconstructs puta-
tive HERV proteins (“puteins”) from the different reading frames in the gene
candidates. The program uses codon statistics, frequency of stop codons and alignments to known retroviral proteins to approximate the original ORF. Tracking of the sequences through the analyses is facilitated by the given identities including genome version, chromosome, and start position.

In the computer assisted investigation of the different available genome downloads, we could build an ERV database and extract sequences for the different subsequent analyses. In papers I-V, we have extracted sequence data from the ERV collection that we subjected to careful analyses. To minimize problems with deleted RT portions, we have found the use of the entire Pol (proteins and “puteins”) suitable for phylogenetic inference of distantly related retroviral sequences (Figure 5).

Figure 5. An unrooted Pol based neighbour joining (NJ) tree with the seven retroviral genera and the more loosely defined ERV class definitions. Host species are indicated with symbols next to each sequence (Adapted from paper V).
Retroviral genera; Pol sequence based classification

**Alpharetroviruses**

Alpharetroviruses, represented by the Avian Leukosis Virus (ALV, figure 5 and Petropoulos, 1997), have only been detected in birds. The history of alpharetroviruses dates to the beginning of the 20th century, when Ellerman and Bang (in 1908), shortly before Rous (in 1911) discovered the first retroviruses (Vogt, 1997a). The alpharetroviruses have a simple genomic setup without accessory genes involved in their replication. During replication, cellular oncogenes can be incorporated into the alpharetroviruses and replace essential structural genes (See Petropoulos, 1997). These viruses become replication deficient and need helper viruses to replicate. An exception is the RSV, which unlike the other alpharetroviruses carry its oncogene (src) in the 3’-end, outside its structural genes (Petropoulos, 1997). In the Pol based phylogenetic tree (Figure 5), alpharetroviruses group close to the betaretroviruses (ERV class II).

**Betaretroviruses**

The betaretroviruses are represented by the mouse mammary tumour virus (MMTV, figure 5 and Petropoulos, 1997). They group, within the ERV class II, as a major branch in the Pol tree (Figure 5). The betaretroviruses have a simple genomic structure with exception for the HERV-K(HML2), which have been shown to code for the accessory proteins Rec or Np9 (Lower et al., 1996; Magin et al., 1999; Armbraester et al., 2002; Armbraester et al., 2004). The betaretroviruses also contain a deoxy-uracil triphosphatase (dUTPase) region in the N-terminal of Pro (Mayer & Meese, 2003) (Paper V). Another distinguishing trait for betaretroviruses is a G-patch in C-terminal Pro (Paper V).

Endogenous betaretroviruses have been subjected to numerous phylogenetic investigations to establish taxonomy. Betaretroviral ERVs have a Lysine-tRNA PBS and are therefore named HERV-K. However, to distinguish groups of elements within the betaretroviruses, similarities to the MMTV has generated the Human MMTV like (HML1-10) classification (Medstrand et al., 1992; Medstrand & Mager, 1998; Andersson et al., 1999; Johnson & Coffin, 1999; Mayer & Meese, 2002; Gifford & Tristem, 2003; Lavie et al., 2004). There are several active betaretroviruses that also have endogenous counterparts e.g. in mouse (MMTV) and sheep (JSRV). The HML2 is the most conserved and also most recently active HERV (paper IV) (Barbulescu et al., 1999; Mayer et al., 1999; Barbulescu et al., 2001; Turner et al., 2001).

Classically, the betaretroviruses have been regarded as mammalian retroviruses found in humans (HML groups), mouse (MMTV), primates (MPMV)
and sheep (JSRV) (Figure 5). However, the newly sequenced chicken genome (International Chicken Genome Sequencing Consortium, 2004) also have detectable betaretroviruses (Figure 5 and paper V). Despite the tight grouping of the betaretroviruses, phylogenetic analyses using Pol and Env have shown that the MPMV (grouping as betaretrovirus in figure 5) is a hybrid element with betaretroviral Pol and a gammaretroviral (ERV class I) Env (Benit et al., 2001).

**Gammaretroviruses**

The gammaretroviruses constitute the largest retroviral genus (Table 4), represented by the murine leukemia virus (MLV, figure 5 and Petropoulos, 1997). All gammaretroviruses have simple genomic structures. The gag, pro and pol are found in the same reading frame and are expressed as a poly-transcript. The Gag-Pro-Pol polyprotein is expressed by a readthrough of a stop codon at the end of gag. This is one of several traits that define gammaretroviruses, extending it from the narrow MLV-like branch (Figure 5) (Papers II and V). The major gammaretrovirus branch is counterbalancing the major betaretrovirus branch mentioned above in the Pol based phylogenetic tree (Figure 5).

Gammaretroviruses can be found in a broad variety of vertebrate species including, mouse, human/primates, but also avian (chicken) and reptiles (Figure 5 and paper V) (Martin et al., 1997). Phylogenetic and evolutionary investigations have been conducted for the largest gammaretroviral (ERV class I) ERVs (Tristem, 2000), HERV-H (Papers I-IV) (Mager & Henthorn, 1984; Hirose et al., 1993; Wilkinson et al., 1993; Lindeskog et al., 1999), and its closely related HERV-F (included into the Adjacent HERV-H like, Paper II) (Kjellman et al., 1999a; Kjellman et al., 1999b; Benit et al., 2003), HERV-W/ERV9 (Voisset et al., 1999; Costas & Naveira, 2000; Voisset et al., 2000; Costas, 2002; Oja et al., 2005) and the closely related HERV-E/HERV-R/ERV3 (Kjellman et al., 1995; Herve et al., 2004; Andersson et al., 2005). In contrast to the betaretroviruses, the most recently active gammaretroviral (ERV class I) ERVs appear in the primates (Paper IV).

**Deltaretroviruses**

The deltaretroviruses, represented by the Bovine Leukemia virus (BLV) and the human T-lymphotropic virus (HTLV) (Figure 5 and Petropoulos, 1997), have complex genomes with the accessory regulatory genes rex and tax. The gag, pro and pol genes are present in different reading frames. The Gag-Pro-Pol transcript requires two successive frameshifts to produce a polyprotein. No endogenous counterparts belonging to the deltaretroviruses have been detected (Paper V and Herniou et al., 1998).
**Epsilonretroviruses**

The Walleye dermal sarcoma virus (WDSV, figure 5 and Petropoulos, 1997) represents the epsilon fish retroviruses. The epsilonretroviruses share the gene structure with the gammaretroviruses with stop codon readthrough after *gag*. However, it has a complex genome structure with three additional ORFs, one upstream of the *gag* and two downstream of the *env*. Also the amphibian retrovirus Xen1 (Kambol et al., 2003) groups together with the epsilonretroviruses (Figure 5). With respect to the genomic structure and translational strategy, the epsilonretroviruses resemble the class I ERVs. However, according to sequences in the Pol tree (Figure 5) they are separate. They branch off earlier than the HERV-H/F (ERV class I) gammaretroviruses, closer to the center of the unrooted Pol tree (Figure 5). Although epsilonretroviruses primarily are found in pisces and amphibians, a few epsilon-like elements were surprisingly detected in the human genome (Oja et al., 2005). This highlights the complexity in retroviral evolution and taxonomy.

**Lentiviruses**

The lentiviruses, represented by the Human immunodeficiency virus (HIV) (Figure 5 and Petropoulos, 1997), have complex genomes with the accessory regulatory genes *rev, tat, vif, vpr, vpu* and *nef*. The *gag* is in one reading frame and the *pro-pol* in another reading frame. The Gag-Pro-Pol requires two successive frameshifts. The non-primate lentiviruses e.g. Equine infectious anemia virus (EIAV) and the Visna virus have a dUTPase, separate from that of betaretroviruses, within the *pol* ORF. HIV is characterized by a strong nucleotide bias enriched for A, probably caused by the cell protein APOBEC3G (as mentioned above and Papers III and V) (Mangeat et al., 2003; Vartanian et al., 2003). No endogenous counterparts belonging to the lentiviruses have been detected (Paper V and Herniou et al., 1998).

**Spuma and Spuma-like retroviruses**

The spuma- and spumalike elements group within the ERV class III (Wilkinson et al., 1994; Lindeskog, 1999; International Human Genome Sequencing Consortium, 2001; Mager & Medstrand, 2003) (Figure 5). The Human foamy virus (HFV) is the prototype for spumaviruses (Petropoulos, 1997). Its genomic structure is complex, with the accessory genes tas (formerly known as *bel1*) and bet (formerly known as *bel2*). The spuma-like elements referred to here includes the HERV-L (Cordonnier et al., 1995; Benit et al., 1999), and related (in Pol) ERVs from both mammals and chickens. The spuma-like elements (ERV class III) are disrupted by mutations acquired under their presumably long (>70 million years) residency in the genomes (Benit et al., 1999). At least the MuERV-L and HERV-L have dUTPases, but in a third position, after the IN, compared to dUTPases in
non-primate lenti (between RT and IN) and betaretroviruses (N-terminal of pro).

**Novel intermediate sequence groups**

Pol phylogeny using ERVs derived from different genomes (Figure 5) shows novel groups between the previously described genera. Between the alpha and betaretroviruses, an intermediate beta-like group derived from chicken ERVs, located together with the reptilian python ERV (PyERV, Huder et al., 2002) (Figure 5, and paper V). The epsilon snakehead retrovirus (SnRV) groups separately from the other epsilonretroviruses in the Pol tree, and are therefore designated an intermediate epsilon-like group.

In summary, the growing wealth of XRV and ERV sequences as products from the sequencing of the human (International Human Genome Sequencing Consortium, 2001), chimpanzee (unpublished data publicly released at [http://genome.ucsc.edu/]), chicken (International Chicken Genome Sequencing Consortium, 2004) and other vertebrate species genomes, will provide the basis for new insights about the vast retrotransposon universe, and eventually an updated taxonomy.

**ERV evolution; forces and consequences**

**Infection and reinfection**

Genomic attacks by retroviral integrations have occurred frequently during the evolution. A screening of the human (version hg16), chimpanzee (version PanTro1) and the chicken (version gg01) genomes, all available for download at [http://genome.ucsc.edu/], demonstrate a bombardment of retroviruses (Tables 3 and 4). Since our screening is rather ERV specific and does not detect all solo LTRs etc., the previous estimation of approximately 8% retrovirus related sequences in the human genome (Bock & Stoye, 2000; International Human Genome Sequencing Consortium, 2001), seems reasonable. The human and chimpanzee genomic difference is approximately 1.2% (Chen et al., 2001; Ebersberger et al., 2002) up to 5% (Britten, 2002), depending on analysis methods. The human and chimpanzee ERV contents are comparable, and the minor differences in table 3 can be assigned to the inferior sequence quality of the chimpanzee draft sequence. This is also true for the chicken genome, which overall had less detectable ERVs, in unity with previous estimations (International Chicken Genome Sequencing Consortium, 2004).
Table 3. RetroTector predicted genomic ERV contents

<table>
<thead>
<tr>
<th>Genome analysed</th>
<th>Total detected elements</th>
<th>Elements with RetroTector© Score&gt;299</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Homo sapiens</em> (hg16)</td>
<td>18213</td>
<td>3164</td>
</tr>
<tr>
<td><em>Pan troglodytes</em> (PanTro1)</td>
<td>13003</td>
<td>2117</td>
</tr>
<tr>
<td><em>Gallus gallus</em> (gg01)</td>
<td>3921</td>
<td>262</td>
</tr>
</tbody>
</table>

1. The compilations are under improvement, depending on sequence draft qualities and RetroTector© optimization.
2. RetroTector© Score >299 confers to true integrations of relatively intact elements.

Based on Pol sequence similarities, the alpharetrovirus genus and the novel alpha-beta group could only be found in the chicken genome, whereas the beta, gamma and spuma-like genera were detected also in the human and chimpanzee genomes (Table 4). The spuma-like genus referred to here includes a diverse and presumably old group of spuma related endogenous retroviral sequences, primarily ERV-L (Cordonnier et al., 1995; Benit et al., 1997; Benit et al., 1999). The gammaretroviruses are the most abundant in all genomes, and the betaretroviruses are the most conserved ERVs in the human and chimpanzee genomes. Alpharetroviruses are the most conserved ERVs in the chicken genome. This is the first presentation of the largest ERV survey conducted to date. However, the compilation in table 4 should be regarded as an overview. A more thorough analysis is in progress. Details will most probably be adjusted slightly with help of new sequence data and improved computer program algorithms (Paper VII, not included in this thesis).

Table 4. RetroTector© defined ERV genera and structures

<table>
<thead>
<tr>
<th>Chromosome:</th>
<th>Alpha</th>
<th>Alpha-Beta</th>
<th>Beta</th>
<th>Gamma</th>
<th>Spuma-like</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome</td>
<td>gg01</td>
<td>gg01</td>
<td>gg01</td>
<td>gg01</td>
<td>gg01</td>
</tr>
<tr>
<td>No. ERVs detected</td>
<td>55</td>
<td>69</td>
<td>628</td>
<td>547</td>
<td>55</td>
</tr>
<tr>
<td>2xLTR in hit</td>
<td>7</td>
<td>5</td>
<td>173</td>
<td>129</td>
<td>5</td>
</tr>
<tr>
<td>detected gag</td>
<td>34</td>
<td>28</td>
<td>369</td>
<td>270</td>
<td>15</td>
</tr>
<tr>
<td>detected pro</td>
<td>30</td>
<td>40</td>
<td>423</td>
<td>341</td>
<td>18</td>
</tr>
<tr>
<td>detected pol</td>
<td>35</td>
<td>45</td>
<td>558</td>
<td>440</td>
<td>25</td>
</tr>
<tr>
<td>detected env</td>
<td>12</td>
<td>2</td>
<td>200</td>
<td>131</td>
<td>0</td>
</tr>
<tr>
<td>Detected &quot;LTR-gag-pro-pol-env-LTR&quot;</td>
<td>4</td>
<td>0</td>
<td>48</td>
<td>29</td>
<td>0</td>
</tr>
</tbody>
</table>

1. A crude compilation of the analysis in progress of RetroTector© detected ERVs with scores over 299.
2. Novel intermediate group as a result of Pol phylogenetic analysis (Figure 5).

The close similarity between the human and chimpanzee genomes (Chen et al., 2001; Britten, 2002; Ebersberger et al., 2002), have made genome-wide comparisons useful tools in interpreting the evolutionary differences. Lineage specific integrations are of utmost interest. Large scale analysis of LTRs have shown integrational differences between humans and apes for the proposed recently active HML2 family (Lebedev et al., 2000; Mamedov et al., 2002), complementing earlier results (Medstrand & Mager, 1998). In humans, the HML2 integrations were shown to have proliferated predominantly via reinfection rather than retrotransposition in cis (Belshaw et al.,
2004). The results were supported by dN/dS (high ds/da) analyses showing overall low values for all HML2 genes including env, thus indicating a negative (purifying) selection. In a following study, it was shown that the reinfection model applied to ERV families with copy numbers less than 200 (Belshaw et al., 2005). The most intact HERV-H elements (926 copies, Paper II) also had low dN/dS, indicating proliferation by reinfection. However, the less intact RTVLH2-like elements (Paper II) did not have significantly low dN/dS, thus indicating another type of proliferation (See “Midwife” elements below and paper II).

Other genomic differences and lineage-specific proliferations have been studied. Recently active elements in the human genome are mainly found in the HML2 family (Paper IV). In primates, on the other hand, gammaretroviral (ERV class I) elements have been active since the human-chimpanzee split, approximately 5 million years ago (Paper IV). Recently a group of ERVs, PtERV1, that could be detected in chimpanzee but not in humans or gorilla, was demonstrated (Yohn et al., 2005). Independently, at the same time, we identified the very same ERV using a bioinformatic approach (Paper IV). Like Yohn et al. we also conclude an exogenous source for the PtERV1/PtNeo-I group (See below and paper IV).

Polymorphism

Most HERV insertional polymorphism studies have been conducted on the betaretroviral HERV-K(HML2) group, probably since these are the most recently active HERVs. The discovery of the almost intact HERV-K(HML2.HOM) (Mayer et al., 1999), increased the interest in both intraspecies and intraindividual comparison of retroviral integrations. Many of the HERV-K elements in the human genome were integrated after the separation of the human lineage from the chimpanzee lineage (Medstrand & Mager, 1998). However, a HERV-K has been reported to have reinfected the germ-line of a common ancestor to humans and apes (gorilla, bonobo and chimpanzee). The apes proved to have an integrated provirus that was absent in humans (Barbulescu et al., 1999; Barbulescu et al., 2001) (See paper IV). Thus the virus has been active for a long time, both before and after the divergence of apes and humans. Insertional polymorphism was reported in humans, where orthologous PCR showed that one provirus HERV-K113(HML2), with full length ORFs, was present in 30% of tested individuals. Another provirus HERV-K115(HML2) was present in 15% (Turner et al., 2001). The integrational polymorphism could also be significant as they provide additional rec loci (See alternative splicing above).

A human polymorphic provirus (HERV-K at chromosome 3q24) with an allele frequency of 0.20 has been reported with flanking inverted repeats
(similar to DNA transposons) proposing a model for an evolutionary link where RNA preceded DNA transposable elements (Hughes & Coffin, 2002). A detailed study using 18 individuals concluded that most HERV-K(HML2) polymorphisms were detected as solitary LTR formations at 5/13 human specific loci (Hughes & Coffin, 2004). The recent activity of HERV-K(HML2) elements in humans is mirrored in several reports of allelic differences between ethnic groups (Macfarlane & Simmonds, 2004; Mamedov et al., 2004). HERV-K113(HML2) has even been suggested as a possible genetic risk marker for some autoimmune and neurodegenerative diseases (See above and Moyes et al., 2005).

Conserved genes, e.g. various ERV $env$ ORFs, have shown a different polymorphism. Single nucleotide polymorphisms (SNP) have been described and analysed in both ERV3 (de Parseval & Heidmann, 1998) and HERV-H (de Parseval et al., 2001). Most of these SNP are synonymous and follow a Hardy Weinberg equilibrium (See HERV-H SNP in paper II), but 1% of the Caucasian population has a non-synonymous SNP in the ERV3 $env$ causing a premature stop and truncated protein (de Parseval & Heidmann, 1998). Variations in HERV-H SU ORF sequences at two loci (chromosome 2 and probably on chromosome 1) were also shown, although with moderately high ds/da ratios (Paper I).

In summary, polymorphisms may have different effects. Insertional polymorphism is an indirect sign for recent transposition or reinfection. The impact of retroviral infection on human evolution is a major question that is difficult to test, partly due to a low transposition/reinfection frequency and partly due to ethical complications in retrieving e.g. brain tissues for the analyses. However, an interesting study was conducted for the L1 retrotransposition activity in brain. The results showed that human L1 generated somatic variation in vitro for rat neuro progenitor cells and in vivo for mice neuronal cells (Muotri et al., 2005). The retrotransposing effects were influencing both gene expression and cell differentiation in the in vitro experiment. Bearing in mind the high expression of ERVs (especially HERV-H) in human brain tissue (Forsman et al., 2005) and recent activity of ERVs (Paper IV), these results are very interesting. A possible connection between retroviral integrations and the human brain development and evolution remains to be shown.

Recombination and gene conversion

Coding regions of an integrated provirus may become lost after a homologous recombination between the two LTRs, leaving a solo LTR at the locus (Stoye, 2001; Hughes & Coffin, 2004). These solitary LTRs are present 10
or maybe 100 times more frequently than their cognate ancestral proviruses (Stoye, 2001).

Recombination can occur in several ways: i. Recombination between LTRs of a single provirus results in a solitary LTR. ii. Homologous recombination between two proviruses on the same chromosome results in loss of viral and genetic sequence between recombination sites. iii. Recombination between 3’ and 5’ LTRs of a given provirus results in a tandem provirus (two proviruses flanked by LTRs while sharing one LTR). iv. Gene conversion resulting in non-homologous gene exchange and no proviral loss. Gene conversion is a recombination event where no crossing over occurs. Two sequences interact in such a way that one is converted by the other (Li, 1997).

Proviral recombinations are not rare events. In a study of HERV-K(HML2.HOM) flanking DNA using sequencing and phylogenetic analysis, it was found that more than 16 % of the proviruses had undergone rearrangement that most probably had led to large chromosomal rearrangements by reshuffling of surrounding regions (Hughes & Coffin, 2001).

Horizontal transfer
Multiple cross-species transfers events have been described for retroviruses along with numerous other microbes (For a review, see Weiss, 2001). Recently the PtERV1 elements (See above and paper IV) were speculated to have multiple origins contributed from chimpanzee together with gorilla, baboon and macaque. Phylogenetic trees of gag and env differed from generally accepted primate species trees, thus indicating horizontal transfers (Yohn et al., 2005). Earlier, BaEV-like viruses were described to have spread among African primates, and probably also to cats, in recent evolutionary time (Figure 6) (van der Kuyl et al., 1995a; van der Kuyl et al., 1995b; van der Kuyl et al., 1996). Indeed, in phylogenetic studies, several vertebrate species show signs of possible transspecies transfers of MLV-like gammaretroviruses (Martin et al., 1997). The SIVcpz (Bailes et al., 2003) and HIV (Gao et al., 1999; Hahn et al., 2000) most likely transferred from smaller primates to chimpanzees, and from chimpanzees to humans, respectively (Figure 6).

Cross-species transmissions may have occurred through various routes. A possible transfer mechanism is exposure of wounds to prey blood during predation (van der Kuyl et al., 1996; Gao et al., 1999; Bailes et al., 2003). Further, integration of retrotransposon DNA after uptake from the alimentary canal cannot be excluded (Forsman et al., 2003).
Master element theories; ERV “Break out” vs. “Midwife”

If mutations of reintegrated TEs are not deleterious they will pass on to the next generation. Thus, mutations accumulate in every retrotransposition and many independent and parallel subfamilies can be formed from an original sequence. Theoretically, such events could lead to an infinite number of mutations over time. A master element, however, can from a single locus produce several new copies which are pseudogenes and cannot multiply themselves. Thus, master element mutations occur in subfamilies formed at different times. The master element theory fits the existing retroelement subfamily data best (Deininger et al., 1992). To study master element evolution from phylogenetic data (e.g. tree topology), simulations have been developed to discern between two alternative models: i. random template model and ii. strict master model. In the random template model, all elements in the genome are equally likely to produce copies, while only a single locus element is capable of this in the strict master model (Clough et al., 1996). Thus, these definitions of possible transposition models promote predictions of TE evolution.
Retrovirus like elements (RLE) may behave somewhat different compared to other TEs as a result of particle formation. The numerous ERV integrations can be considered as “proviral quasispecies” which are not static. Occasionally they change by gene conversion and/or homologous recombination (Johnson & Coffin, 1999; de Parseval et al., 2001; Hughes & Coffin, 2004). An RLE quasispecies with a gag ORF could in theory provide a functional capsid and another RLE quasispecies with env ORF could provide the envelope proteins (Figure 7). Copackaging of the two different quasispecies into particles could occur and result in recombination during reverse transcription. If damaged regions are omitted by the recombination events, the virus can infect a new cell, albeit with lower efficiency. The process is self-optimizing and is generating a virus that “breaks out” (Blomberg et al., 2005). A related process has been reported as “patch repair” in MLV (Mikkelsen et al., 1996). These events may explain the occurrence of particles and cause trouble in relating them to phylogenetic data.

The disrupted retrovirus can under appropriate circumstances be complemented in trans as suggested (Mager & Freeman, 1995). A large number of HERV-H-like elements (RTVLH2-like in paper II) are deficient in Pol (RT4 and 5 (Shih et al., 1989)) but still have more similar LTRs, than a much smaller RGH2-like subgroup of HERV-H with more complete Pol (Paper II). These findings (Mager & Freeman, 1995 and paper II) indicate a proliferation where the deficient RTVLH2-like elements were “delivered” by the RGH2-like elements, hence called “Midwife” elements. The deficient RTVLH2-like elements were thus given a new chance to reintegrate, through complementation in trans by their “midwives”. Recently, in a study of synonymous and non-synonymous substitutions, the “Midwife” model gained further support. The RGH2-like elements (more complete but old according to the LTRs, paper II) with a low dN/dS ratio applied to a reinfection model and the RTVLH2-like elements (less complete but younger according to the LTRs, paper II) indicated a complementation in trans (Belshaw et al., 2005). It was speculated that due to low probability, these complementations will only be found in large ERV families. Interestingly the small group of HERV-Fc elements (Benit et al., 2003), now included into the larger Adjacent HERV-H like group (Paper II), possibly have midwife properties. The single copy number element HERV-Fc1 (AL354685) is, despite its divergent LTRs (5.7% nonidentity), almost intact in gag and pol, and intact in pro and env. It may have provided functional proteins to many defective HERV-H-related elements during the evolution of higher primates while, for unknown reasons, it could not transpose itself.
Figure 7. The “ERV Break out” hypothesis. Two proviral quasispecies with deleterious mutations indicated by “X” in different genes. Co-packaging and jumping polymerase, theoretically lead to a self optimization and increased retroviral fitness. (Modified from Blomberg et al., 2005).
The present investigation

The large HERV-H/F like group

The large HERV-H family (also named RTVL-H and RGH) have been estimated to about 100 full length elements, about 1000 deleted sequences (adjusted to 926 in paper II), and about 1000 solitary LTRs (Mager & Henthorn, 1984; Hirose et al., 1993; Wilkinson et al., 1993). Although low copy numbers of HERV-H have been detected in new world monkeys (NWM), which diverged from old world monkeys (OWM) about 40 million years ago (Mager & Freeman, 1995), the majority of HERV-H integrated in the genome of OWM about 30-35 million years ago (Mager & Freeman, 1995; Anderssen et al., 1997). The great expansion of HERV-H integrations have been estimated to occur before the evolutionary split between OWM and Apes, about 30 million years ago (Goodchild et al., 1993).

Despite a long time residency of HERV-H in the genome, and accumulation of numerous deleterious mutations, it still has env ORFs (Lindeskog et al., 1999). In the human genome there are three loci with env ORF (Paper I and de Parseval et al., 2001). The MLV TM, similar to the HERV-H TM, has been observed to have immunosuppressive properties (See the ISU/CKS17 motif in figure 2) (de Parseval et al., 2001; Mangeney et al., 2001). These immunosuppressive properties of MLV TM were recently demonstrated to be required in vivo for melanoma tumour growth in mice (Mangeney et al., 2005). Expression of HERV-H like sequences (also including the former HERV-F group, cf. paper II) have been observed in normal leukocytes (Lindeskog et al., 1998) and in leukemia (Patzke et al., 2002). This also makes it intriguing that the HERV-H env with its ISU still has an ORF. Screenings for HERV-H like ESTs also increase the interest in the cancer connection. HERV-H showed higher expression in stomach, colon, prostate, and testis tumour compared to other HERVs (Paper III and Stauffer et al., 2004).
The genomic significance has been further established in studies that have shown the use of HERV-H as alternative promoters for native genes. The alternative splicing from the major SD to a gene downstream of the provirus is used to fuse the start in 5’LTR with downstream native genetic information (Feuchter & Mager, 1990; Feuchter-Murthy et al., 1993; Kowalski et al., 1999). The alternative splicing within HERV-H (Figure 8 and appendix B) (Wilkinson et al., 1990; Lindeskog & Blomberg, 1997) indicates complexity in genome structure, normally a feature not associated with gammaretroviruses but has been observed in betaretroviruses (See alternative splicing above).

The HERV-H like elements are divided into the RTVLH2-like and the RGH2-like subgroups, based on Pol similarities (Paper II). RGH2-like elements have accumulated more mutations in their LTRs, and are therefore considered older than the RTVLH2-like elements. Proliferation of the Pol deficient RTVLH2-like elements is probably assisted by complementation in trans (Mager & Freeman, 1995) facilitated by more complete “Midwife” elements (Paper II). A probable “Midwife” element may be the single copy HERV-Fc1 (Benit et al., 2003), included into the Adjacent HERV-H like group (Paper II).

Thus the contributions of the large HERV-H/F group are significant for understanding the gammaretroviral (ERV class I) evolution. An alternative method to study the evolutionary progress is “retro-evolution” and in silico reconstruction of putative ancestral retroviruses based on residing ERVs (Paper III). The putative ancestral HERV-H provirus has a PBS, complementary to the human histidine-tRNA (Paper III and Mager & Henthorn, 1984). It is approximately 9kb and maintains a genomic structure similar to that of the gammaretrovirus prototype MLV (Figure 8, appendix B and paper III).
Figure 8. Schematic overview of the annotated putative ancestral HERV-H provirus consensus ORFs. Three reading frames and alternative splicing in the HERV-H consensus sequence as interpreted by RetroTector© and by EST searches. The proviral genes with outlined names of conserved consensus motifs are presented as black bars below each reading frame. Alternative splice patterns are outlined with their corresponding splice donor (SD) and splice acceptor (SA) sites (Adapted from paper III).
Aims of the study

The aims of the present study were to:

- Investigate the amount and variation of integrated HERV-H proviruses in the human genome.
- Analyse the large HERV-H like group and study its evolutionary relations within the group and compared to other retroviruses.
- Characterize a putative ancestrall HERV-H, using in silico “retro-evolution”, for comparisons of similarities and differences against other retroviruses.
- Examine the differences in recent endogenizations between the human and chimpanzee genomes in an attempt to establish evolutionary links.
- Use the wealth of XRV and ERV sequences, made available through the genome sequencing projects, to establish structural markers that can be projected over sequence phylogenies and thereby infer an updated retroviral taxonomy.

The genetic influences described above have made it important to study and characterize the retroviral content of the human genome, and compare it against other species, to establish functionally important properties. An in silico bioinformatic approach for sequence analysis is a natural succession of the sequencing projects for the human (International Human Genome Sequencing Consortium, 2001), chimpanzee (available publicly at [http://genome.ucsc.edu/]) and chicken genomes (International Chicken Genome Sequencing Consortium, 2004). Phylogenetic analyses (Johnson & Coffin, 1999) are indispensable tools to dissect the results and specific ERV characteristics may be subjected to further expression and functionality studies.

Paper I

Using polymerase chain reaction (PCR) primers derived from HERV-H19, we amplified, cloned and sequenced HERV-H SU from a normal human blood sample. Amplified sequences were used for searching similar sequences with or without ORFs in the nonredundant (nr) and high-throughput ge-
nome sequencing (htgs) databases made available through GenBank (www.ncbi.nlm.nih.gov). ORF-containing sequences clustered with HERV-H sequences from chromosomes 1 and 2. The 5' and 3' LTRs, from SU ORF- and non-SU ORF-containing elements, differed about 4 % and indicated a similar time of integration. The sequences had a moderately high number of synonymous vs. non-synonymous mutations, which indicated a possible selection for maintenance of the HERV-H SU ORFs. Two variants of HERV-H SU were found in the blood sample and reported to GenBank (accession numbers: AY05287 and AY050298). Together with three other env ORFs (AJ289710 HERV-H/env60 and AJ289711 HERV-H/env59) (de Parseval et al., 2001), we presented evidence for SU polymorphism and selection, however with low ds/da ratios, for ORF function in two HERV-H loci, on chromosome 2 and probably on chromosome 1.

Paper II

Using a newly developed bioinformatic tool RetroTector© (Paper VII, not included in this thesis), we screened the human genome for HERV-H like integrations. The data obtained were used in phylogenetic analyses to define the HERV-H group and relate it to other gammaretroviruses. We found 926 HERV-H elements among 1124 detected HERV-H like out of a total of 3661 detected pol containing ERVs (Table 5). After applying an iterative alignment procedure on the entire Pol and constructing several phylogenetic trees, we were able to divide the "HERV-H like" elements into 926 HERV-H with two subgroups: "RTVLH2-like" (705 elements) and "RGH2-like" (77 elements). The remaining sequences were designated the "Adjacent HERV-H like" group. Among 926 HERV-H, LTR differences were 1–33%, 10% had env, 78% had gag, 66% had a histidine primer binding site (PBS), and 3% (both subgroups) had a phenylalanine PBS (Table 5).

Table 5. Summary of HERV-H like sequences in the human genome.

<table>
<thead>
<tr>
<th>HERVH-like</th>
<th>HERV-H</th>
<th>Adjacent HERVH-like</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>&gt;80% nt identity to:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RTVLH2</td>
</tr>
<tr>
<td>Integrations</td>
<td>1124</td>
<td>926</td>
</tr>
<tr>
<td>5' and 3'LTRs</td>
<td>395</td>
<td>320</td>
</tr>
<tr>
<td>H-PBS</td>
<td>669</td>
<td>610</td>
</tr>
<tr>
<td>F-PBS</td>
<td>105</td>
<td>26</td>
</tr>
<tr>
<td>gag</td>
<td>858</td>
<td>727</td>
</tr>
<tr>
<td>pro</td>
<td>479</td>
<td>393</td>
</tr>
<tr>
<td>pol</td>
<td>1041</td>
<td>926</td>
</tr>
<tr>
<td>env</td>
<td>129</td>
<td>91</td>
</tr>
</tbody>
</table>
The analysis of PBS types gave no evidence for a separate HERV-F group, albeit a tendency for more phenylalanine PBSes in the Adjacent HERV-H like group. Defects in Pol RT were observed. They may cause trouble in creating broadly amplifying PCR methods, as many HERV-H lack part of the most frequently targeted motif (...YXDD...) (Xiong & Eickbush, 1988; Shih et al., 1989).

The findings of more complete pol in proviruses with older and more mutated LTRs, made us formulate a "Midwife" master model where more complete elements help the more disrupted elements in trans to transpose. Complementation by a more complete "Midwife" element in trans could take place via capsids (MA and CA), encapsidation (NC), the retroviral enzymes (PR, RT, RH, and IN), and envelope (SU and TM). A more efficient packaging of RNA from the older and more defective elements than of the "Midwife" element could help to mobilize older more deleted elements to new integrations, in which their LTRs would start out identical again. Recently, this phenomenon was speculated to occur only in large ERV groups due to a low probability for complementation in trans (Benit et al., 2003). The small HERV-Fc group, now included into the larger Adjacent HERV-H like group, have the expected "Midwife" properties. A likely "Midwife" candidate could be the single copy HERV-Fc1 (AL354685, included into the Adjacent HERV-H like group) that despite its divergent LTRs (5.7% nonidentity), is almost intact in gag and pol, and intact in pro and env. The difference compared to the ERV "Break out" hypothesis (See above and Blomberg et al., 2005), is that the "Midwife" model does not involve co-packaging of retroviral RNA and recombination to produce more replication prone offspring.

We also analysed allelic differences in env at chromosome 2q24.3 (See paper I) using site specific PCR and a temperature gradient gel electrophoresis method (TGGE). A known non-synonymous single nucleotide substitution in SU (de Parseval et al., 2001) was used as a method control and in analysis of Swedish blood donors. We found the substitution to follow a Hardy-Weinberg equilibrium and after comparison to the results of de Parseval et al., we suggested a selection neutrality for homo- and heterozygotes in this position in the European population.

Paper III

We used the HERV-H sequence information obtained from genome-wide screenings (Paper II), to construct a putative ancestral HERV-H provirus in an in silico "retro-evolutionary" approach. In alignment based comparisons of 18 representative HERV-H sequences (widely distributed in the HERV-H group in paper II) and prior retroviral knowledge, we applied minor manual
adjustments to the primary consensus and created a full length HERV-H ORF sequence (Figure 8 and appendix B). The general structure resembled the prototypic gammaretrovirus, MLV (Petropoulos, 1997).

Unlike other hitherto described gammaretroviruses, the reconstructed HERV-H Gag had two NC zinc fingers (Figure 9). This structural trait may become useful in a revision of the gammaretroviral evolution. The distantly related viruses of the major betaretroviral branch (See figure 5) and the errantiviruses have two NC zinc fingers (Figure 9). Thus, the two zinc fingers probably existed early in gammaretroviral evolution. The second NC zinc finger, detectable in the HERV-H elements, appears to have been lost among the MLV-like gammaretroviruses. We imagine that the lost zinc finger gradually became compensated by a charged amino acid section (Cheslock et al., 2003) upstream of the remaining zinc finger. The charged amino acid section (most developed in MLV, figure 9) locates in a region overlapped by the Gag interaction (I) domain, which is required for virion formation (Bowzard et al., 1998).

Figure 9. Alignment of Gag CA-NC. The first box indicates the MHR region (MHR consensus derived from Benit et al., 1997). The second box indicates the charged amino acid segment (marked with *-consensus) upstream of the zinc fingers (CCHC motifs, boxes 3 and 4) (Adapted from paper III).
Additionally, we found an unusually long “pre-gag” sequence ranging from the PBS to the normal methionine start of gag (Figure 8 and appendix B). The gag ORF could be extended in the 5’ end, similarly to a feature described in MLV that produces a glyco-Gag (Fan et al., 1983). The function of the HERV-H elongated “pre-gag”, is however obscure. Alternative splicing patterns described above (and in figure 8) were confirmed by EST data where env transcripts could be detected in various cancer tissues and were coherent with the results of others (Patzke et al., 2002; Stauffer et al., 2004).

Interestingly, the HERV-H had a pronounced nucleotide frequency bias, with over 29% C and below 17% G (Appendix A). The same bias, however less pronounced, could also be detected in the deltaretroviruses (Berkhout et al., 2002). It is reasonable to assume that HERV-H-like sequences, like lentiviruses, also met an innate antiretroviral defence involving a host RNA editing enzyme. However, the mechanism is unknown and must be different from the cytidine deamination caused in HIV by the cellular APOBEC3G. A possibility that the bias is caused by a selectively error prone RT cannot be excluded (Katz & Skalka, 1990; Vartanian et al., 1994; Berkhout et al., 2001).

Paper IV

The Human-Chimpanzee speciation has been estimated to have occurred between 4.6 and 6.2 million years ago. The genomic difference is approximately 1.2% (Chen et al., 2001; Ebersberger et al., 2002) up to 5% (Britten, 2002), depending on analysis methods. Using the RetroTector© program (Paper VII, not included in this thesis), we screened the two genomes in a search for differences in ERV integrations. We analysed ERVs with LTR similarities over 98%. With 0.2% mutations per million year (Li, 1997), the 2% LTR difference served as an approximation to 5 million years and the human and chimpanzee split.

We found 51 pol containing “recent” ERVs that were included into a phylogenetic tree (Figure 10). Using stringent criteria including a control against ERV flanking sequences in the next genome, 30 of the 51 sequences were found to be unique to respective genome. The sequences segregated into four groups. In the human genome, they segregated into a betaretroviral group, HERV-K(HML2), and the gammaretroviral HERV-H like group (Figure 10). Besides these two groups, there were also two additional gammaretroviral groups named PtNeo-I (Ia and Ib) and PtNeo-II (IIa and IIb) in the chimpanzee which did not occur in the human genome.
In chimpanzee, the recent unique integrations were dominated by 27 gamma-retroviral sequences, but also one betaretroviral (Figure 10). The groups were defined using the criterion of >80% Pol sequence similarity. The recently described PtERV1 (Yohn et al., 2005), clustered together with our PtNeo-Ia subgroup.

The PtNeo groups were most similar to a colobus, a macaque and two baboon ERVs, but neither to other chimpanzee nor to any human gammaretroviruses. The pattern was consistent with cross-species transfer via predation (van der Kuyl et al., 1996; Gao et al., 1999; Bailes et al., 2003) and applies...
to the larger cross-species transmission network in figure 6. Thus, several trans-species transfers of non-human, non-chimpanzee primate gammaretroviruses to chimpanzee occurred since the human-chimpanzee split.

The comparison of ERVs in the human and chimpanzee genomes highlights the importance of habitat, interspecies contact, predator-prey relations and cross species retroviral infection, and/or stochastic reactivation of pre-existing ERVs, as determinants of the retroviral genetic setup of a species.

**Paper V**

In an attempt to investigate possible structural markers that are useful for projection over sequence based phylogenetic trees and thus strengthen the taxonomy, ERVs with relatively complete structures were retrieved from the genetic archives of humans and chickens, and analysed, using RetroTector© (Paper VII, not included in this thesis). The genome wide screenings demonstrated a variety of proviral sequences but none belonging to the delta or lentiviruses, thus confirming previous analyses (Herniou et al., 1998). Over 3300 proviral sequences were collected into a database, from which we could extract representative sequences for further analyses. Twelve representative proviral sequences were used in addition to previously described ERVs and XRVs to reconstruct a broad Pol phylogenetic tree (Figure 5). The large data collection enabled us to analyse ERVs of previously not described intermediate groups between the recognized genera.

Retroviral classification using host species is at first sight appealing: Classical gammaretroviruses are murine, epsilon piscine, alpha avian and beta mammalian. However, in figure 5, this order does not apply when additional ERVs are included into the analysis. The human and chicken genomes both have ERVs clustering in five retroviral genera.

A presumably original translation strategy with all essential protein coding genes in a single poly-ORF, is present in the possibly still active Cer1 gypsy/Ty3 class retroelement, integrated into chromosome III in the *Caenorhabditis elegans* genome (Britten, 1995). The gypsy elements are only distantly related to the retroviruses (Figure 5). The translational strategy divided the tree in figure 5 into two major branches. One contains the gammaretroviruses (ERVclass I) together with the epsilonretroviruses, and another which includes betaretroviruses (including class II ERVs) together with delta, lenti and alpharetroviruses with their respective intermediate groups. The schematized tree, with projections of structural markers, is presented in figure 11.
The modular C terminal IN GPY/F-motif is an especially interesting marker. It is present in a major part of the schematized tree, except in the alpha, beta and their intermediate sequences (Figure 11). To the GPY/F another "chromo" (chromatin-binding) domain is sometimes appended, which interacts with chromatin via DNA-binding proteins (Malik & Eickbush, 1999; Sandmeyer, 2003). Alterations in this domain may alter the specificity of the integration (Singleton & Levin, 2002). This may be relevant for the recently described, however different, nucleotide preferences for HIV, ALV and MLV integration sites (Holman & Coffin, 2005). It was speculated that mechanisms other than primary sequence recognition were involved since an absolute consensus for the integration site of each retrovirus was missing. Thus, a possible involvement of the GPY/F domain in the selection of integration sites should be further investigated.

The G-Patch (Aravind & Koonin, 1999; Gifford & Tristem, 2003; Svec et al., 2004) is a very specific marker for delineation of betaretroviruses, where it can be detected in the C-terminal Pro. The intermediate beta-like ERVs have like the betaretroviruses a dUTPasePro, but are distinguished by the lack of a detectable G-Patch (Figure 11).

![Figure 11](image-url). Summary of structural markers that are useful to project over sequence phylogenetic trees (See figure 5) in order to conclude a stronger taxonomy. The number of NC zinc fingers, presence of dUTPase (although not of spuma-like viruses), known accessory genes, C-terminal Pro (G-patch) and Pol (GPY/F) motifs are shown. Nucleotide bias was defined to 25±5 %, (↑) shifted upwards; (↓) shifted downwards; (≡) uncertain bias. Exploration of the LTR lengths of the different groups as detected by RetroTector© are shown as boxplots. In addition, the translational strategy may be used in the phylogeny to separate the gammaretroviruses (including ERVclass I) from spuma-like elements (ERVclass III), deltaretroviruses, lentiviruses, alpharetroviruses and the betaretroviruses (ERVclass II) with respective intermediate groups. The Gypsy and Copia are not included in the translational strategy analysis (Adapted from paper V).
In a supplementary phylogenetic analysis using 389 dUTPase sequences, we got evidence for three separate acquisition events (not shown here). The three dUTPases could further be detected in different retroviral genomic positions. The betaretroviral dUTPase \textsuperscript{Pro} was located N-terminal of \textit{pro}. The non-primate lentiviral dUTPase \textsuperscript{PolA} was located between RT and IN, whereas the dUTPase \textsuperscript{PolB} in the spumalike ERV-L was located C-terminal of IN (See retroviral classification above). The human betaretroviral dUTPase \textsuperscript{Pro} and chicken beta-like retroviral dUTPase \textsuperscript{Pro} formed one branch together with the more studied mammalian betaretroviral MMTV and MPMV dUTPase \textsuperscript{Pro} sequences. This indicated a monophyletic origin of dUTPase \textsuperscript{Pro}. Judging from the phylogenetic tree, acquisition of dUTPase \textsuperscript{PolA} (by non-primate lentiviruses) and dUTPase \textsuperscript{PolB} (by the spumalike ERV-L) may also have been single events.

Pol similarities and other structural markers that are suggested to be useful in taxonomy (Figure 11) are either “global”, applicable to differentiation of distantly related sequences, or “local” for delineation of more closely related sequences and delineation of groups.

“Global” genomic properties useful in phylogenies are:
- Translational strategy.
- Number of NC zinc finger motifs (See also paper III).
- Presence of Pro N-terminal dUTPase (dUTPase \textsuperscript{Pro}) (Baldo & McClure, 1999; Mayer & Meese, 2003).
- Presence of Pro C-terminal G-patch (Aravind & Koonin, 1999; Gifford & Tristem, 2003; Svec et al., 2004).
- Presence of a GPY/F motif in the Pol integrase (IN) C-terminal domain (Malik & Eickbush, 1999).

“Local” retroviral genomic properties useful for delineation of groups are:
- Host species range (Figure 5).
- Nucleotide compositional bias (Appendix A).
- LTR lengths.
Concluding remarks and future prospects

The human genome is littered with ancient retroviral integrations. The HERV-H is one of the largest gammaretroviral (ERV class I) groups. With its 926 members, it has become a prime target for both laboratory analyses and in silico bioinformatic approaches. Possible connections to various diseases have encouraged studies of the HERV-H characteristics and evolution. The present investigation has provided new insights in the large HERV-H group, regarding its sequence variation, structural characteristics, and evolutionary relation to other retroviruses.

HERV sequence variations (Papers I and II) and insertional polymorphisms show that some ERVs are not fixed in the genomes. Reintegrations may still occur. The large HERV-H group show general gammaretroviral characteristics when compared to the prototypic MLV. Since HERV-H branches off early in the Pol phylogenetic tree (See figure 5), it is an interesting model for gammaretroviral evolution. Unique characteristics are e.g. two NC zinc fingers, a pronounced nucleotide composition bias and “Midwife” master element properties. A plausible gammaretroviral “Midwife” ERV is the single copy HERV-Fc1 in the Adjacent HERV-like group that may complement more disrupted ERVs in trans. The high expression of HERV-H in brain tissue (Forsman et al., 2005) brings the reintegration topic forward. Possible similarities with the LINE’s influences on cell proliferation and differentiation should be investigated. Differences in ERV activity between species may contribute to a better understanding of this issue. Ideally the ERV content of the red jungle fowl (International Chicken Genome Sequencing Consortium, 2004) should be compared to the genome of the domesticated white leghorn chicken.

Horizontal cross species transfers exemplified by the lentiviral HIV (Gao et al., 1999; Hahn et al., 2000) and now also indicated for the gammaretroviruses in chimpanzee (Paper III and Yohn et al., 2005), are additional events that add to the genomic variations. An analysis of domesticated animals may reveal if this is an ongoing process. The study should not be limited to predator-prey transmissions as discussed in paper IV. The dog genome should be suitable for comparison against the human genome, because of a close coexistence of the two species in recent evolutionary time. The domesticated ungulates (cow and sheep) should be compared against each other for com-
mon recent integrations. The betaretrovirus JSRV may be a suitable model with its residing ERV counterparts in sheep. A comparison against the human genome may reveal transfers via the alimentary tract (Forsman et al., 2003) or exposure to blood in similarity with the predator-prey transmissions discussed for the chimpanzee in paper IV. Since these possible integrations confer to recent time, they are likely to be polymorphic in the population. Experimental analyses using PCR and sequencing should therefore be implemented. Otherwise the integrational polymorphism will be undetected.

An alternative hypothesis to the recent active, or horizontally transmitted, ERVs observed in the various genomes is the awakening, self optimization and ultimately a “Break out” of residing ERVs (Blomberg et al., 2005). To distinguish between the different modes of recent retroviral activity, integration loci should be analysed in detail with respect to the retroviral genomic structure and compared to orthologous sites in related species. Integration site specificity and possible involvement of appended IN segments downstream of the GPY/F motifs also needs attention. ERV ages estimated by LTR differences may be compromised by homologous recombination, homogenization by gene conversion and variable mutation rates within the host genome.

A number of structural markers were useful for phylogenetic inference. Projection of them over sequence phylogenetic trees increases the strength in the taxonomic and evolutionary conclusions. To improve the conclusions further, sequences derived from non-Pol retroviral genomic portions should be used in parallel to avoid artefacts by recombination. Protein sequences allow a more flexible comparison and are therefore favoured over nucleotide sequences when comparing distantly related ERVs and XRVs.

To understand and fight the challenges brought to humans by retroviral infections, we must learn to understand the basic retroviral concepts. Inference of phylogeny and an improved taxonomy helps us to find common denominators for the retroviruses. Addition of ERVs, made possible through numerous sequencing projects, brings new insights into retroviral evolution. This thesis has, with a focus on the large gammaretroviral (ERV class I) HERV-H, provided and dealt with the study of ERV evolution.
Acknowledgements

I thank the Uppsala University and the Department of Medical Sciences, Section of Virology for a stimulating research environment. I express my sincere gratitude to:

My ever enthusiastic supervisor, Jonas Blomberg, for introducing me to the fascinating field of endogenous retroviruses, supporting the scientific work and sharing his great knowledge in numerous interesting discussions, not only concerning our profession.

My co-authors and in particular Göran Sperber for fruitful collaboration and sharp minded computer programming. I thank Ann-Catrin Andersson, Erik Larsson, Göran Andersson, Mats Lindeskog, Patrik Medstrand, Hong Yin, Raymond (Awad) Josephsson and all other “retrovirologists” for valuable discussions.

All the present and former members of the Virology Research group and the Clinical Virology for creating an interesting environment. Working in the “United Nations of Virology” has really broadened my insight into foreign cultures. I thank all the present and former members of the Bacteriology Research groups. Björn Herrman, Sandra Hjalmarsson and Caroline Fock for initial lab introductions and support. I also credit all the people at Clinical Bacteriology and the Infection control unit for their patience and sharing of localities.

Bengt Kallin, Ann Catrin Andersson and Eva Haxton for careful reading and comments on this thesis.

All the helpful people around, family and friends that managed to endure my non-stop “preaching” and ventilation of ERV thoughts. Mats Pettersson, Kalle Näslund, Tobias Norström, Tobias Backström, and Jonas Persson for almost daily lunch meetings through all the study years, staying updated, and for the traditional crawfishing every autumn.

Finally, I express my appreciation to Stefan and Judit Modin and Elisabeth Bölenius for all the nice moments that truly have been re-energizing.
References


Brosius, J. (1999). Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* 107, 209-38.


Muster, T., Waltenberger, A., Grassauer, A., Hirschl, S., Cautig, P., Romirer, I., Fodinger, D., Seppele, H., Schanab, O., Magin-Lachmann, C., Lower, R.,


Appendices
Appendix A. Retroviral (XRV and ERV) nucleotide frequencies

<table>
<thead>
<tr>
<th>Genus</th>
<th>Virus</th>
<th>A</th>
<th>T(U)</th>
<th>G</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>NC_001407.1</td>
<td>Rous sarcoma virus, complete genome</td>
<td>24</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>NC_001408.1</td>
<td>Avian myelocytomatosis virus, complete genome</td>
<td>23</td>
<td>19</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>NC_001408.1</td>
<td>Avian sarcoma virus, complete genome</td>
<td>24</td>
<td>23</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>NC_001401.1</td>
<td>Fujinami sarcoma virus, complete genome</td>
<td>22</td>
<td>18</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>NC_001402.1</td>
<td>Avian sarcoma virus, complete genome</td>
<td>24</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>Beta</td>
<td>NC_001503.1</td>
<td>Mouse mammary tumor virus, complete genome</td>
<td>30</td>
<td>26</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>NC_001550.1</td>
<td>Mason-Pfizer monkey virus, complete genome</td>
<td>30</td>
<td>27</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>NC_004994.2</td>
<td>Enzootic nasal tumour virus of goats, complete genome</td>
<td>28</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>NC_004994.1</td>
<td>Ovine pulmonary adenocarcinoma virus, complete genome</td>
<td>29</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>MMTV M15122</td>
<td>30</td>
<td>26</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>HERV-K/HML2(11, 105040013)</td>
<td>32</td>
<td>26</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>HERV-V-MM14</td>
<td>31</td>
<td>27</td>
<td>19</td>
<td>24</td>
</tr>
<tr>
<td>Delta</td>
<td>NC_001815.1</td>
<td>Simian T-lymphotropic virus 2, complete genome</td>
<td>23</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>NC_000858.1</td>
<td>Simian T-lymphotropic virus 1, complete genome</td>
<td>23</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>NC_001436.1</td>
<td>Human T-lymphotropic virus 1, complete genome</td>
<td>23</td>
<td>23</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>NC_003232.1</td>
<td>Primata T-lymphotropic virus 1, complete genome</td>
<td>24</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>NC_001444.1</td>
<td>Bovine leukemia virus, complete genome</td>
<td>22</td>
<td>24</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>HERV-K</td>
<td>25</td>
<td>29</td>
<td>15</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>HERV-K like : chr1</td>
<td>30</td>
<td>28</td>
<td>17</td>
<td>31</td>
</tr>
<tr>
<td>Gamma</td>
<td>NC_001702.1</td>
<td>Murine type C retrovirus, complete genome</td>
<td>25</td>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>NC_001363.1</td>
<td>Murine sarcoma virus, complete genome</td>
<td>25</td>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>NC_001362.1</td>
<td>Friend murine leukemia virus, complete genome</td>
<td>25</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>NC_001501.1</td>
<td>Murine leukemia virus, complete genome</td>
<td>26</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>NC_001502.1</td>
<td>Moloney murine sarcoma virus, complete genome</td>
<td>25</td>
<td>21</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>FeLV NC_001940</td>
<td>28</td>
<td>22</td>
<td>23</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>HERV-H consensus</td>
<td>24</td>
<td>28</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>HERV-H like : chr1</td>
<td>30</td>
<td>28</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>HERV-H like : chr11</td>
<td>30</td>
<td>28</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>HERV-H like : chr11</td>
<td>30</td>
<td>28</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>HERV-H like : chr3</td>
<td>30</td>
<td>28</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>HERV-H like : chr5</td>
<td>30</td>
<td>28</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>HERV-H like : chr5</td>
<td>30</td>
<td>28</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>HERV-H like : chr5</td>
<td>30</td>
<td>28</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>HERV-H like : chr5</td>
<td>30</td>
<td>28</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>HERV-H like : chr5</td>
<td>30</td>
<td>28</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>HERV-H like : chr5</td>
<td>30</td>
<td>28</td>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>HERV-H like : chr5</td>
<td>30</td>
<td>28</td>
<td>15</td>
<td>35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Selected groups (no. elements)</th>
<th>A (avg/SD)</th>
<th>T(U) (avg/SD)</th>
<th>G (avg/SD)</th>
<th>C (avg/SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>24,22,20,25</td>
<td>15,5/1,5</td>
<td>30,7/2,3</td>
<td>30,7/2,3</td>
</tr>
<tr>
<td>Adjacent HERV-H like (556)</td>
<td>26,6/3,3</td>
<td>17,0/2,2</td>
<td>29,1/3,9</td>
<td>29,1/3,9</td>
</tr>
<tr>
<td>ERV9 and HERV-E like (383)</td>
<td>29,9/1,8</td>
<td>21,5/1,5</td>
<td>23,7/1,7</td>
<td>23,7/1,7</td>
</tr>
<tr>
<td>HIV and HERV-E like (169)</td>
<td>29,9/1,7</td>
<td>24,3/1,6</td>
<td>23,7/1,7</td>
<td>23,7/1,7</td>
</tr>
<tr>
<td>XRV</td>
<td>29,8/1,7</td>
<td>20,1/1,6</td>
<td>20,1/1,6</td>
<td>20,1/1,6</td>
</tr>
<tr>
<td>XRV-Elike (674)</td>
<td>29,8/1,7</td>
<td>20,1/1,6</td>
<td>20,1/1,6</td>
<td>20,1/1,6</td>
</tr>
</tbody>
</table>
Appendix B. Putative ancestral HERV-H sequence
Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 62

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)