Università degli Studi di Cagliari

DOTTORATO DI RICERCA

Biologia e Biochimica dell'Uomo e dell'Ambiente

Ciclo XXVI

Characterization of Human Endogenous Retrovirus sequences identified in the human genome using the RetroTector software

Settore scientifico disciplinare di afferenza

Microbiologia BIO/19

Presentata da:	Laura Vargiu
Coordinatore Dottorato	Prof. Emanuele Sanna
Tutor/Relatore	Prof. Enzo Tramontano

Esame finale anno accademico 2012 – 2013

Acknowledgements

I would like to express my deep gratitude to my supervisors, Prof. Enzo Tramontano from the University of Cagliari and Prof. Jonas Blomberg from the University of Uppsala, for their understanding and support.

A special thanks to Prof. Jonas Blomberg for accepting me for three months in his laboratory in Uppsala and introducing me to the world of Human Endogenous Retroviruses. He gave me the opportunity to actively collaborate with him in the HERV classification, a huge task achieved through the application of his newly developed Simage methodology. His continuous advice on my thesis project during my stay in Uppsala and during this last year have been priceless. I really appreciate his teachings and each of those exchanges we had on HERV. I know how much I owe him in this thesis.

I also would like to thank all my dear colleagues of the DS3 group at CRS4. Especially Dr Patricia Rodriguez-Tomé, the head of the group, for providing a good environment for both work and enjoyment during these last three years.

Finally, I owe special gratitude to my family and close friends for their continuous and unconditional support of all my undertakings, scholastic and otherwise.

Abstract

Human Endogenous Retroviruses (HERVs) represent the inheritance of ancient germ-line cell infections by exogenous retroviruses and the subsequent transmission of the proviral integrated elements to the descendants. Actually, no replication-competent HERV sequence is recognizable in the human genome. However, some HERVs retain one or several intact retroviral genes and may express protein products that could interfere with the human immune system. The number and classification of HERVs vary according to method of enumeration. The focus of this project is to perform a systematic analysis and a classification of the most intact HERV sequences in order to better understand their evolution and their involvement in shaping the human genome.

The human genome assembly GRCh37/hg19 was analyzed with RetroTector software and a total of 3290 HERV proviral sequences were identified.

The complex genetic structure of the 3290 proviruses was resolved through a multi-step classification procedure that involved a novel type of similarity image analysis (Simage). The 3290 HERVs were classified in 40 unique clades (groups) which could be placed into class I (Gamma- and Epsilon-like), II (Beta-like) and III (Spuma-like). Simage analysis contributed to define the presence of a high number (around 40%) of mosaic forms, with heterogenous sequence content.

A finest characterization of the HERV sequences was achieved with the investigation of a broad panel of structural markers that contributed to confirm and extend the previously performed classification.

Finally, the HERVs background of integration was also studied. Integration patterns analysis showed a tendency for proviruses from the same clade to occur together, within 100000 bases, maybe due to local duplications. Representatives from some gammaretroviral clades (HERVH and HERVE) integrated more frequently than expected by chance into the 5´end of transcriptional units, mostly in antisense orientation. A few lncRNAs were also found to contain HERV sequences. Thus, *cis*-effects from HERVs are to be expected.

In conclusion, this study represents an advance in the state-of-the art of

HERVs characterization within the human genome and a starting point for upcoming studies on HERVs.

Table of contents

1. Introduction

1.1 Retroviruses

Retroviruses are a large family of animal and human pathogen RNA viruses that share specific features, such as the structure, the composition and the replication strategy. The peculiarity of Retroviruses is that, after virus attachment and penetration inside the host cell, the retroviral life-cycle (Fig. 1.1) involves two distinctive steps mediated by two specific viral enzymes, the reverse transcriptase (RT) and the integrase (IN), respectively: i) the reverse transcription of the RNA viral genome into a linear double-stranded DNA and ii) the subsequent integration of this DNA into the genome of the host cell (Goff, 2007).

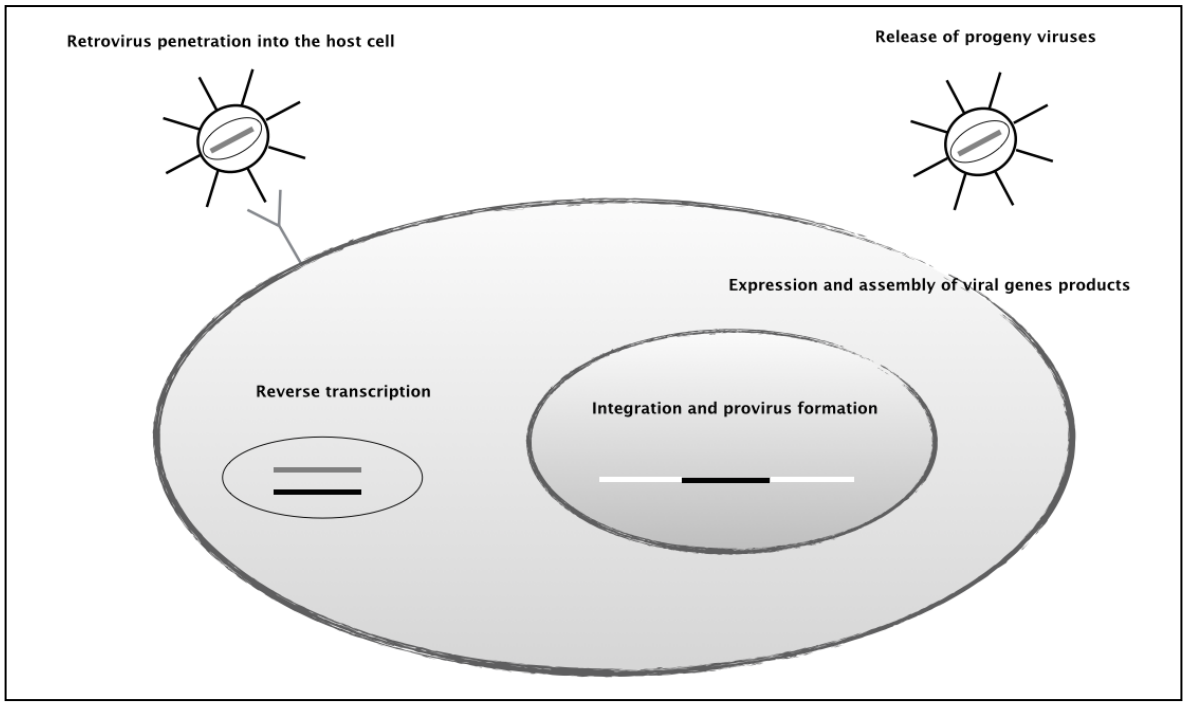

Fig. 1.1 A simple view of the retrovirus life-cycle.

The integrated form of the viral genome, named provirus, is genetically stable (integration is essentially an irreversible step) and can be thus considered as a "cellular gene". Therefore, the provirus can be regularly transcribed, throughout the host translational machinery, giving rise both to new copies of the RNA viral genome and to the structural viral proteins necessary to assemble the viral progeny that can be released from the host cell.

The specificity of Retroviruses replication strategy is responsible for their pathogenic impact and their role in various diseases such as tumors. Indeed, the integration process can be viewed as an insertional mutagenesis and a mechanism that can contribute to oncogene activation and cellular transformation (Nevins JR, 2007).

Based on their genome organization (Fig. 1.2, from Vogt, 1997), Retroviruses can be defined as simple or complex. The simple RNA viral genome of Avian leukosis virus (ALV) has the following features: i) a 5' cap structure and a 3' poly(A) tail; ii) a short repeated (R) sequence present both at 5'- and 3'-ends; iii) a U5 (unique) sequence immediately downstream the R region located at 5'-end; iv) a U3 (unique) sequence immediately upstream the R region located at 3'-end and v) an internal core sequence coding for the main viral proteins (Gag, Pro, Pol and Env). In contrast, the complex retroviruses, such as Human T-lymphotropic virus (HTLV), have additional sequence information for the production of small regulatory proteins with different functions (Tax and Rex).

The U5 and U3 regions of the viral genome contain regulatory sequences necessary both for the provirus integration (att sites) and for the regulation of the viral genes expression (promoter, enhancers and poly-A signals). During the reverse transcription step, the viral genomes are subjected to structure and size re-organization. In fact, both the U3 and U5 regions are duplicated and translocated, leading to the formation of long terminal repeats (LTR) that is a ‘U3-R-U5’ block sequence flanking the resulting double-stranded DNA. The generated LTRs harbor all the required regulatory sequences necessary to drive not only the integration process but also the provirus transcription, since promoters located in the 5’LTRs are particularly efficient in starting (triggering) the transcription process (Goff, 2007). The integrated provirus start to be transcribed by host-cell RNA polymerase II from the promoter located on the U3 region of the 5’ LTR up to the transcription stop signal (poly-A) located at the U5 sequence of the 3’-terminal LTR. The transcription process generates full-length viral RNA transcripts that serve both as new viral genomes and as mRNAs that are processed (spliced) to obtain the structural (Gag and Env) and

enzymatic (Pol) viral proteins. Depending on retrovirus genome complexity, the splicing patterns vary from one (simple retroviruses) to multiple (complex) giving rise to the production of both the Gag-Pro-Pol-Env as well as the accessory proteins (i.e. Tax and Rex).

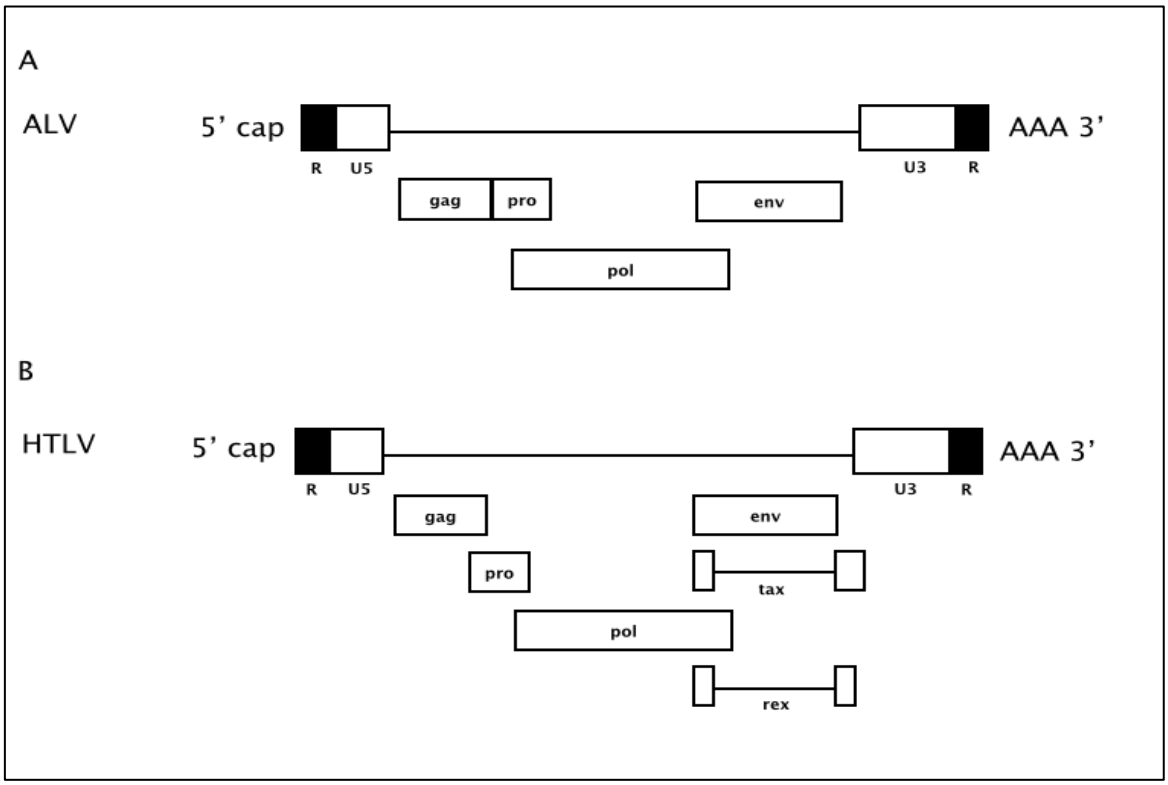

Fig. 1.2 Examples of Retroviruses genome organization. A) ALV (Avian leukosis virus) has a simple retroviral genome with both the 5' and 3' terminal noncoding sequences (R, U5 and U3) that encompass the coding regions for the main four viral proteins (Gag, Pro, Pol and Env); B) HTLV (Human T-lymphotropic virus) has a complex structure with additional coding sequences for accessory (regulatory) proteins (Tax and Rex) (modified from Vogt, 1997).

The current taxonomy of Retroviruses is based on the 9th (2012) report released by the International Committee on Taxonomy of Viruses (ICTV). Actually, all known retroviruses are included within the Retroviridae family and can be further divided in two subfamilies that embrace a total of 7 different genera (Table 1.1).

Generally, Retroviruses are known to infect somatic cells and can be horizontally transmitted within members of a host population. Thus implying that, once the infection is established, the provirus can be eradicated only if all the host cells carrying the viral genome are

eliminated, that is the "natural" death of the host. However, when germ-line cells are infected by Retroviruses, the viral genome integration give rise to an endogenization process with a consequent (resulting) vertical transmission of proviruses (according to Mendelian laws) to the host off-spring. Therefore, the provirus can be inherited generation by generation and can be fixed within the host population.

Table 1.1. Classification of Retroviridae family

[a]Subfamily	[a]Genus	Species	[b]Genome
Orthoretrovirinae	Alpharetrovirus	Avian leukosis virus (ALV) Rous sarcoma virus (RSV)	simple
	Betaretrovirus	Mouse mammary tumor virus (MMTV) Mason-Pfizer monkey virus (MPMV) Jaagsiekte sheep retrovirus (JSRV)	simple
	Deltaretrovirus	Bovine leukemia virus (BLV) Primate T-lymphotropic virus type 1 (HTLV-1)	complex
	Epsilonretrovirus	Walleye dermal sarcoma virus (WDSV)	complex
	Gammaretrovirus	Murine leukemia virus (MLV) Moloney murine sarcoma virus (MoMuLV) Feline leukemia virus (FeLV)	simple
	Lentivirus	Human immunodeficiency virus type 1, 2 (HIV-1, HIV-2)	complex
Spumaretrovirinae	Spumaretrovirus	Simian foamy virus (SFV)	complex

[a]International Committee on Taxonomy of Viruses (ICTV) 2012; [b]Retroviruses genome organization (see details in the text).

First observations of retroviruses endogenization process date back to 1950s when evidences of retroviral genomes were found in uninfected cells of chicken and mice. Since then, endogenous retroviruses (ERV) have

been found in all vertebrates, including humans (Boeke & Stoye, 1997). In some cases, retroviruses can co-exist both as exogenous and endogenous forms in their host populations (i.e. MMTV, JSRV and KoERV) (Tarlinton et al., 2006; Baillie et al., 2004) but, most endogenized viruses represent a "relic" of ancestral exogenous retroviruses infections, such as human endogenous retroviruses (HERVs).

1.2 What are HERVs?

Since the first "draft" of the human genome sequence has been available (Lander et al., 2001), it was clear that transposable elements and retroelements accounted for almost half of the human genome. In particular, retroelements represent a broad group of sequences that could amplify within the genome through a RNA intermediate (Deininger & Batzer, 2002).

As far as their general classification, retroelements are divided in two groups, the non LTR- and the LTR-elements, based on the absence/presence of the LTR (Fig. 1.3, modified from Balada et al., 2009). The first and more abundant (around 34% share of the human genome) group encompasses both short and long interspersed repetitive sequences (SINEs and LINEs, respectively) which possess a limited protein coding capacity, so that SINEs elements rely on the reverse transcriptase codified by LINEs to amplify and retrotranspose themselves in new genomic positions.

The second group that accounts for around 8% of the human genome comprises a wide variety of retroviral-like element, such as retrotransposons and human endogenous retroviruses (HERVs). As shown in Fig. 1.3, all these sequences resemble the genetic organization of exogenous retroviruses, seemingly differing from each other for the presence of one of the four typical retroviral genes, the Env gene. However, the borderline between retrotransposons and HERVs is not so strictly defined, as most HERV sequences are devoid of a recognizable Env gene whereas some retrotransposons (Ty3/gypsy elements) have a

distinguishable Env gene (Malik, 2000).

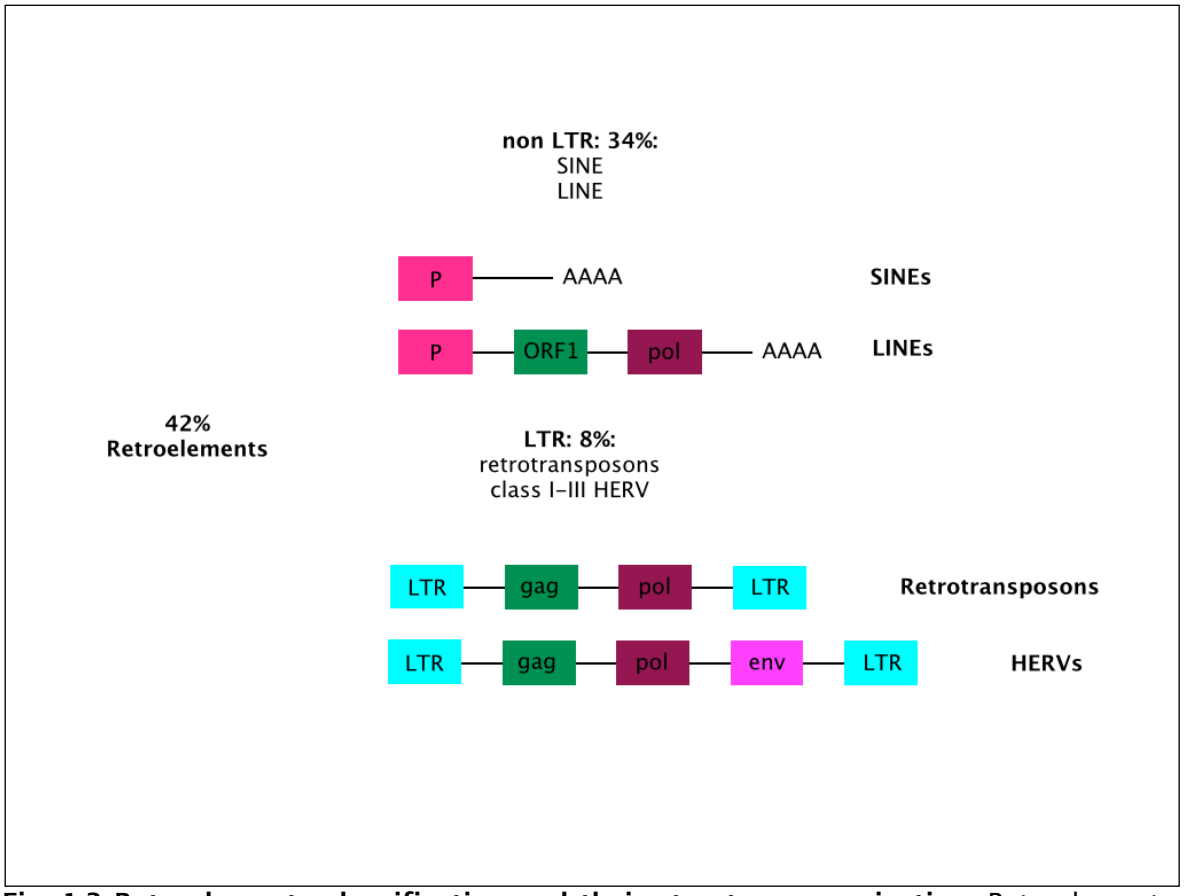

Fig. 1.3 Retroelements classification and their structure organization. Retroelements are distinguished in: non LTR-element (SINEs: short interspersed elements; LINEs: long interspersed elements and pseudogenes) and LTR-elements (HERVs: human endogenous retroviruses and other retrotransposon elements). The given percentages represent the relative proportions to the human genome (Deininger & Batzer, 2002) (modified from Balada et al., 2009).

The phylogenetic evolution of the HERV sequences showed that most of them entered the mammals and primate genomes over 30 million years ago (Bannert & Kurth, 2006; Katzourakis et al., 2005b). Since the first integration waves, most HERVs have been severely damaged in their original genetic structure by accumulation of mutations, insertions and deletions up to the total excision of the internal coding region through homologous recombination between the two flanking LTRs (Stoye, 2012; Copeland et al., 1983; Benachenhou et al., 2009). The solo-LTRs formation has contributed to increase the number of HERV traces within the human genome, since in most cases the presence of an ancient integrated provirus can be assessed by the identification of a single-LTR and its

specific integration sites. Except for the more recently integrated group of human species-specific HERVK(HML2) and a few other HERVs that still retain some coding protein potential, the existence of a replication competent retroviral sequence capable to produce an infectious virus particle has not yet been demonstrated. Nonetheless, the presence of this great amount of genetic material of retroviral origin may impact the human genome at various levels modulating and shaping its physiological functions (Lower, 1996).

1.3 State-of-the-art of HERV classification

An important issue of HERV research regards the different methodologies that have been applied for the identification and classification of the retroviral sequences. Wet-lab and bioinformatics/computational approaches were both used to define the actual enumeration of HERV sequences, both proviral (most integer sequences) and solo-LTRs.

Generally, HERV were identified and classified according to sequence similarity, mainly using the Pol gene as a marker of phylogenetic inference, with their exogenous counterparts (Boeke & Stoye, 1997; Andersson et al., 1999; Tristem, 2000; Katzourakis & Tristem, 2005a). This approach have led to a number of identified HERV groups (also improperly named as "families") ranging between 26 to 31 with the possibility of further increase. Moreover, the copy number of sequences within each group varied from a few sequences (i.e. the HERVFC) up to the large HERVH "family" that has been predicted to account for, at least, almost 1000 more or less intact members and an equal number of solo-LTRs derived from HERVH internal sequence homologous recombination. Until now, the definitive list of HERV groups present in the human genome and their copy number is not definitively assessed.

However, according to the above described classification methods, HERV groups can be broadly divided in three main classes: i) class I HERV related to Gammaretroviruses; ii) class II HERV related to Betaretroviruses and iii) class III HERV related to Spumaretroviruses.

Another important issue is on the HERV nomenclature that it is still not standardized. Historically, HERV names are linked to the different approaches/methodologies applied for their identification leading to a puzzle of names sometimes difficult to interpret and translate. The need for the introduction of a definitive and standard HERV nomenclature has been recently introduced (Blomberg et al., 2009; Mayer et al., 2011).

1.4 Impact of HERV on human genome

The conservation of HERVs within the human DNA over the time can be regarded as a balance between beneficial and detrimental effects that could be exerted by these sequences towards the host organism. HERVs can alter/modulate the human genome physiology through three different mechanisms: 1) direct expression of viral proteins and/or RNA transcripts; 2) interference and control of the host transcriptome and 3) modulation of the genome plasticity through the insertion of bulk DNA (Stoye, 2012).

Generally, the presence of HERV RNA transcripts as well as the expression of viral protein or retroviral-like particles (RVLPs) have been detected both in healthy and pathological human tissues and, in many cases, these findings have been tentatively associated to different human diseases (Voisset et al., 2008 and reference herein).

Among the best-studied HERV proteins are syncytin-1 and -2 because of their role in placenta formation within primates. Syncytin-1 is the product of the Env gene expressed by one HERV-W provirus. This protein has been detected in the cells that form the external layer of the placenta (Màllet et al., 2004) and its fusogenic property is involved in placental syncytiotrophoblast formation. Besides, the Env gene product of HERV-FRD provirus has been identified and characterized as a fusogenic protein sharing similar function with syncytin-1 and it has therefore termed syncytin-2 (Blaise et al., 2003). These latter represent two well-defined examples of positive selection (co-option) of retroviral genes in participating to physiological development process because of their beneficial effect for the host. Apart from the positive role in placenta

development, syncytin-1 has also been associated to some diseases, such as breast cancer (Bjerregaard et al., 2006) and multiple sclerosis (MS), a chronic inflammatory disease of the central nervous system characterized by tissue inflammation and demyelination. Indeed, it has been observed that MS patients showed an overexpression of syncytin-1 in astrocytes (neural cells) where it induced free radicals formation which damaged the myelin producing oligodendrocytes leading to demyelination (Antony et al., 2011).

Another negative side of HERV proteins is represented by Rec and Np9, both expressed by the HERVK group, which may have a role in germ cell tumors development (seminomas and teratocarcinomas) (Armbruester et al., 2004; Galli et al., 2005). In fact, it has been demonstrated that both these proteins (Denne et al., 2007) can interact with a transcriptional repressor protein, the promyelocytic leukemia zinc finger (PLZF), a crucial factor for leukemogenesis in humans and spermatogenesis in mice. The observed increased frequency of teratocarcinoma, associated with a not properly functioning PLZF, led to the conclusion that the interaction between Rec and Np9 and PLZF could be responsible for tumor development.

A most broad HERV impact on the human trancriptome is represented by the influence that these retroviral sequences could have if they integrate in proximity of human genes. In particular, as mentioned above, the HERV LTRs, either associated to - more or less intact - proviruses or solo-LTRs, represent a wide source of promoters (alternate or bidirectional), enhancers, repressors (poly-A signals) or alternative splicing sites for human gene transcripts (Jern & Coffin, 2008; Leib-Mösch et al., 2005; Medstrand et al., 2001). A striking example of how HERV can affect the human gene context of integration is represented by the human salivary amylase cluster where the integration of HERVE sequences promotes a tissue-specific expression of the gene in the parotid glands. Other examples are represented by the pleiotrophin (PTN) gene expression in placenta that is mediated by another HERVE element or the leptin obesity hormone receptor (LEPR) that exists in two forms generated by an

alternative splicing due to a HERVK LTR.

Due to the fact that exogenous retroviruses are known to be pathogenic in animals, many efforts have been involved to find an analogous correlation between HERV and different human diseases, such as cancer, multiple sclerosis and autoimmune diseases (Balada & Ordi-Ros, 2010; Brodziak et al., 2012; Cegolon et al., 2013).

However, the general opinion is that a definitive proof of HERV-induced disease and a complete picture of how and where HERV act is still far away to be demonstrated (Young et al., 2013).

1.5 Retrotector

The bioinformatics and computational approach is one of the most used in retroelement and LTR-retrotransposon identification and it is confirmed by the increasing number of developed software (Lerat, 2010).

RetroTector is a program package (Sperber et al., 2007) implemented for the identification of endogenous retroviruses integrated in vertebrate genomes, including those of primates and humans. Briefly, the main RetroTector framework relies on a complex set of algorithms that, overall, leads to a complete reconstruction of the retroviral-integrated sequences (see also Chapter 2 and Chapter 3).

RetroTector has many advantages, such as the possibility to identify full integrations, not only bits or short sequence pieces, the reconstruction of retroviral protein (putein), the estimation of open reading frame (ORF) and a preliminary retroviral genus classification. Moreover, RetroTector is not dependent on repetition (a priori provirus detection), thus implying the capacity to identify low-copy number retroviral sequences, i.e. HERVFc.

However, RetroTector is not fully optimized for a complete identification of some class III sequences, such as Spumaretrovirus-like and mammalian apparent LTR-retrotransposon elements (MaLR), as well as single-LTR detection.

Since its development, RetroTector has been successfully used for the detection of ERV/HERV in more than 30 different vertebrate genomes (Fig. 1.5, from Blikstad et al., 2008), generating a wide Retrobank collection of endogeous sequences useful for retroviruses evolution studies (Blomberg et al., 2010).

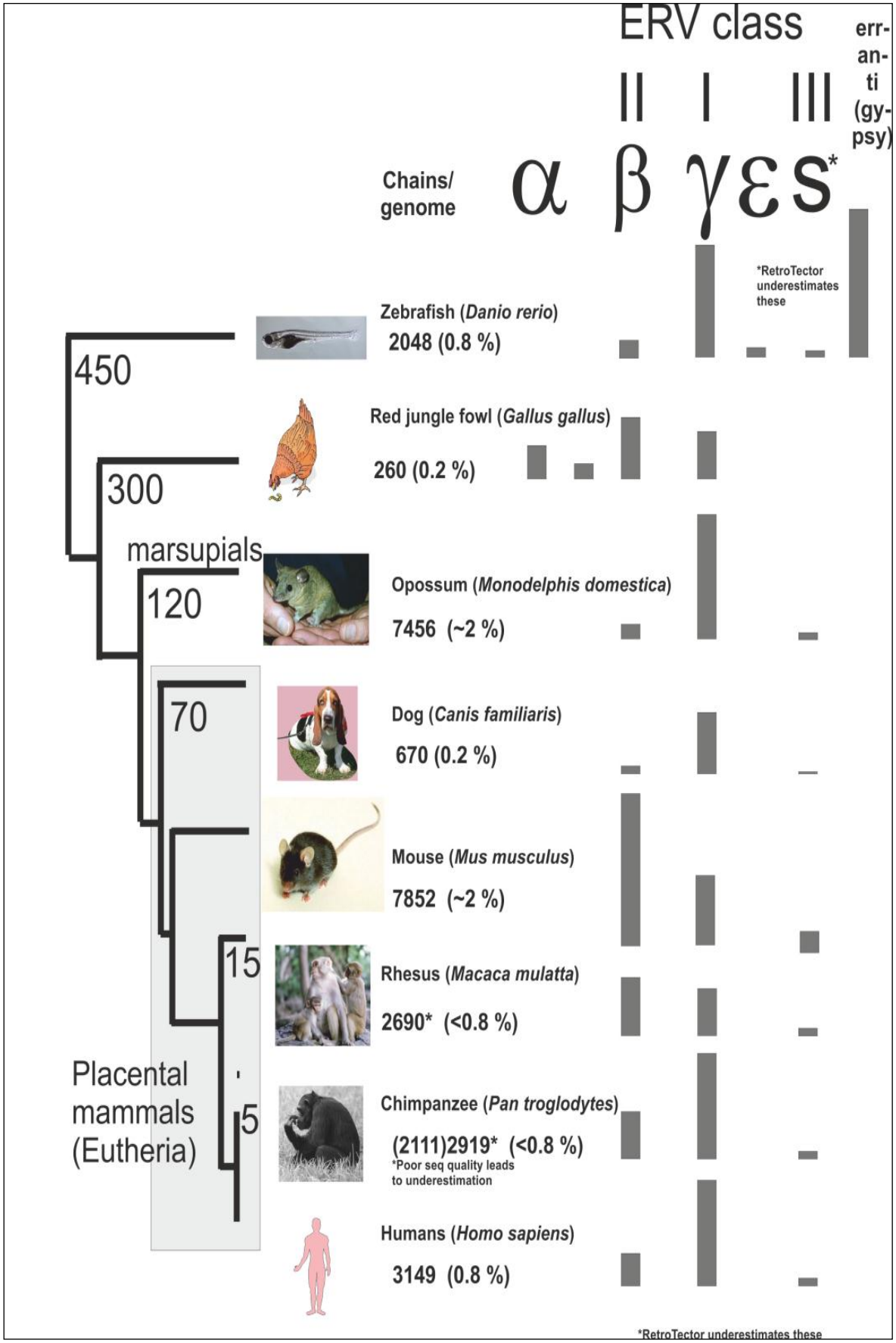

Fig. 1.5 Overview of Class I, II and III ERVs. Phylogenetic analysis of ERVs identified with RetroTector in eight vertebrates including human (from Blikstad V. et al., 2008).

1.6 Aim of this project

The above described issues concerning the up-to-date HERV classification and the time-consuming and sometimes difficult retrieval of publicly available retroviral sequences could led to miss a complete overview of the HERV world.

Thus, the focus of this project was firstly to identify and, secondly, to perform a complete, as far as possible, characterization of the most intact HERV sequences (solo-LTRs were not considered) found within the human genome.

In order to address these objectives, a bioinformatics approach with the RetroTector software, for the identification of HERV proviruses in the human genome GRCh37/hg19 was preferred.

The final aim is to get a 360° overview of the Who and Where of HERV sequences and contribute to improve the knowledge about their evolution and their involvement in shaping the human genome.

2. Materials and Methods

2.1 Human genome assembly (GRCh37/hg19)

The February 2009 assembly GRCh37/hg19, released by the Genome Reference Consortium (Lander et al., 2001), is the human reference sequence used to perform the HERV identification.

The full haploid set (22 + X +Y) of chromosomes sequences was downloaded, as FASTA files (chr*.fa.gz), *via* the UCSC Genome browser (http://genome.ucsc.edu/) and the file storage was set up at the CRS4 Institute on an Intel based machine.

2.2 Retroviral consensus and reference sequences

The different data sets of retroviral consensus and reference sequences, used to perform the HERV multistep classification procedure (see Chapter 3), were obtained as follows:

i) an exhaustive data set of both exogenous and endogenous retroviral sequences (**RvRef**) was collected by Jonas Blomberg from the literature data with the principle of precedence for the first publication of the sequence;

ii) a set of 67 consensus sequences (**RvRefx**) was generated through the joining of HERV chains into groups using an *in house* tree-based clustering algorithm (“ConsensusFromTree") (Blomberg J., unpublished);

iii) a set of 9 HML (**HML**) consensus sequences, generated for the HERVK (HML1-HML9) group, were kindly provided for this project by Vidar Blikstad and Jonas Blomberg (unpublished data);

iv) the entire Repbase Update (**RU**) (Jurka et al., 2005), a database of repetitive DNA elements was downloaded at:

http://www.girinst.org/repbase/update/index.html;

v) the “LTR” subset from the entire Repeatmasker (**RM**) collection of vertebrate repeats (release of May 2012) (Smit, 1996) was downloaded at: http://www.repeatmasker.org.

2.3 RetroTector

The human genome GRCh37/hg19 was examined for the presence of HERV proviral sequences using the RetroTector software (version 1.01), a program package developed for the recognition of endogenous retroviral sequences in vertebrates genomes (Sperber et al., 2007).

RetroTector is mainly based on the principle of "fragment threading", an algorithm that searches the whole genome for the presence of conserved motif hits (from known exogenous and endogenous retroviral proteins) and their distance constraints. If these parameters are satisfied, RetroTector attempts to reconstruct the proviral sequences (namely chains), to predict the putative retroviral proteins (referred to as puteins) sequences and to estimate the original longest ORF (open reading frame) for each putein. A preliminar classification of the identified proviruses (chains), based on a RetroTector viral genus assignment, and a chainscore that identifies the degreee of the chains intactness are also given. The data generated during the analysis are stored in a database and the results could be visualized through a user-friendly interface and extracted, as table format, for detailed investigations.

RetroTector was set up at the CRS4 Institute on a computing cluster, an Intel based machine with 4 Xeon processors with 6 2.66 GHz cores, 256 Gb of RAM with an estimated execution time for the GRCh37/hg19 of 1-2 days.

2.4 HERVs classification tools

The classification of the identified HERV chains was performed as a multistep procedure based on Pol amino acid and whole chain nucleotide

similarity to different collections of retroviral reference and consensus sequences (**RvRef, HML, RU and RM**).

The MEGA software (version 5.2) (Tamura et al., 2011) was used for sequence alignment and phylogenetic trees inference. Multiple alignments were performed using both Muscle and ClustalW with default settings. The neighbor-joining trees were based both on Pol amino acid and nucleotide sequences, and bootstrap analysis was carried out with 1000 replicates.

A custom algorithm, "ConsensusFromTree" (Jonas Blomberg, unpublished) was used to generate a set of consensus sequences (**RvRefx**) by grouping HERV chains according both to Pol amino acid and whole chain nucleotide sequence similarity. The degree of heterogeneity for each group was calculated and the consensus sequences were used for successive phylogenetic inference.

The final HERV classification was accomplished with the creation of Simage (similarity image) (Jonas Blomberg, unpublished). Each HERV chain was sliced into twentieths, and the similarity blast score (graded from 0 to 9) of each twentieth to the entire collections of retroviral reference and consensus sequences (RvRef, RvRefx, HML and RM) was determined. The number of positions in a target twentieth that matched the search sequence was used by Blastpars2 program to generate the simage score with the maximum of similarity (all positions matched) set to 9. The other values (from 9 to 0) were derived by comparison of the number of matching positions to the total number of positions in the given twentieth.

Simage where more than half of the best matching twentieths derived from the same reference sequence, and less than four twentieths were “0”, that means a total absence of similarity to a reference sequence or to a closely related reference sequence, were considered unambiguous (or canonical) representatives of the most frequently matching reference sequence. In cases where both RvRef and RM indicated an unambiguous reference sequence, preference was given to the RvRef sequence.

A final set of 40 HERV consensus sequences represented by the chains

with the most intact Gag, Pro, Pol and Env ORFs (the “best representatives”) was obtained. The latter were generated through ClustalW alignments of both whole nucleotide chains and puteins (Gag, Pro, Pol and Env) within each HERV classified group (clade). The degree of heterogeneity of the groups, that is the portion of positions identical in more than 50% of members (heterogeneity index), the portion of gaps in the alignment, and the average of both "intermember identity within the group" (IWIG) and "identity to consensus within the group" (ITC) were calculated.

2.5 PBS quality control

The "Homo sapiens" subset of tRNA sequences, used to perform a comparative quality control of the HERV primer binding sequences (PBS) sequences identified by RetroTector, was downloaded from the tRNA database (Jühling et al., 2009) at http://trna.bioinf.uni-leipzig.de/DataOutput/. A comprehensive (BLAST) search for matches, with up to two mismatches, in the first 1000 nucleotides of the 3290 HERV chains, identified by RetroTector, was made.

2.6 HERV clusters of integration

HERV clusters of chromosomal integrations were calculated by slicing chromosomes in 10 million-base bins and mapping back the chromosomal coordinates of the HERV sequences.

A HERV cluster was defined when more than 20 endogenous proviral integrations occurred within a 10 million-base bin considering that the random distribution of the 3290 HERV sequences within the 3.2Gb of GRCh37/hg19 would generate 1 provirus every $9x10^5$ bp.

2.7 HERV sequences integrations with respect to Transcriptional Units

A transcriptional unit (TUx) was defined as a nucleotide stretch of the

human genome that could be transcribed into a RNA molecule, regardless of its protein coding potential. The annotated data sets of TUx for the GRCh37/hg19 were obtained as follows:

i) Ensembl 71 and Vega 51, the data sets of protein coding genes were downloaded at:

http://www.ensembl.org/index.html;

ii) the GENCODE data set (version 7), a collection of 14880 manually curated and annotated long non-coding RNA (lnc-RNA) (Derrien et al., 2012; Harrow et al., 2012) was dowloaded at:

http://www.gencodegenes.org/data.html.

HERV integration pattern analyses with respect to TUx data sets were made by custom algorithms (kindly performed by Jonas Blomberg). Each chromosome was divided into bins (fragments) and the minimum distances from the HERV chromosomal coordinates within the start and end positions of the TUx and in the range of 0-1, 1-10, 10-20 up to >90 kb (kilobases) both upstream and downstream the TUx were calculated.

3. HERVs identification and classification

3.1 Introduction

As described in Chapter 1, the aim of this project was to perform a systematic analysis and the subsequent classification of the most intact HERV sequences that integrated in the human genome during evolution.

In order to perform such analysis, the human genome assembly GRCh37/hg19 was screened using the RetroTector software (Sperber et al., 2007), a program package developed for the recognition of retroviral sequences in entire vertebrate genomes.

Each chromosome of the GRCh37/hg19 was analyzed, according to a flow of operations that has been highly automated. The files of the vast amount of data generated during the analysis could be visualized as a table with the total number of the identified and reconstructed retroviral sequences, designated by RetroTector as chains, and for each chain, a list of parameters (score, chain-genus, sub-genes, start, end and others) that contribute to the overall identification of the retroviral sequences.

Before proceeding to the complete classification and analysis of the identified HERV sequences, a first elaboration of the output data was needed due to the fact that RetroTector sometimes produces alternative interpretations of the same chain. Therefore, overlapping and/or alternative chains (with similar score or similar start and end position in a chromosome) or chains with a low chain-score (old and very fragmented elements) were removed, both *via* algorithms and manual inspection. Next, some of the RetroTector parameters such as the chain-genus and a collection of reference and consensus sequences, obtained from the literature (the “RvRef” collection of nucleotide and amino acid sequences) or from appropriate Internet sites (Repbase), were utilized to perform a first elaboration of the retroviral chains.

It is well known that the taxonomy and the nomenclature of HERV is still a matter of debate (Blomberg et al., 2009; Mayer et al., 2011). In fact, it is

based on nucleotide and amino acid similarity to exogenous retroviruses but recombination events, secondary integrations and deletions during evolution have contributed to the actual composition of the HERV and often lead to difficulties in classification due to the ambiguity of the sequences. It is worth to note, that the definition of a "definitive" HERV nomenclature, clearly linked to HERV classification, is out of the scope of the present study. The present efforts, in fact, have been mainly directed to classify the "Who" and "Where" of the HERV found in the human genome.

3.2 Raw data and first elaboration of HERV sequences

RetroTector was run on the GRCh37/hg19 and a total number of 5547 proviral chains were identified as RetroTector output showing a chain-score range of 250-4000. This great amount of data was further filtered as follows:

1) according to an established procedure (Sperber et al., 2007), a chain-score cutoff 300 was applied to remove most of the proviruses that could be considered to be artifacts as well as to avoid the oldest and most incomplete chains. With this procedure the number of chains was hence reduced roughly by 30%;

2) the remaining 3703 chains were then scrutinized for the presence of duplicated/overlapping sequences. These were recognizable by checking out the "Start" and "End" positions (but also to CoreFirst and CoreLast parameters). When duplicated chains occurred, the one with the highest chain-score was preferred. This inspection was performed firstly *via* a specific algorithm, secondly the removed overlapping chains were manually inspected to avoid missing some chains that could be real tandem duplications.

Overall, the results of this data mining gave a definitive number of 3290 HERV sequences together with a preliminary and rough viral classification (Table 3.1). The preliminary classification inherent to RetroTector was based on Pol amino acid and nucleotide similarities of the detected retroviral chains compared to different collections of both exogenous

retroviruses and HERV reference sequences obtained from i) literature data (RvRef), ii) Repbase Update (RU) a database of repetitive DNA elements (Jurka et al., 2005) and iii) a set of 9 HML (HML) consensus sequences elaborated by Blikstad et al. (unpublished).

Table 3.1. General identification and preliminary classification of human endogenous retroviruses (HERV) in GRCh37/hg19

[a]Genera	Species	[b]HERV genus	[c]N° of chains	[d]N° of clades
Gammaretroviruses	Murine leukemia virus (MLV) Moloney murine sarcoma virus (MoMuLV) Feline leukemia virus (FeLV)	Gamma-like	1291	21
Betaretroviruses	Mouse mammary tumor virus (MMTV) Mason-Pfizer monkey virus (MPMV) Jaagsiekte sheep retrovirus (JSRV)	Beta-like	516	10
Spumaretroviruses	Simian foamy virus (SFV)	Spuma-like	175	2
		Unclassified	1308	-
		Total	3290	33

[a]Retroviruses classification from International Committee on Taxonomy of Viruses (ICTV) 2012; [b]HERV genus determined by RetroTector (score cutoff 300); [c]number of HERV chains (sequences) identified by RetroTector; [d]number of provisional classified HERV clades (groups).

The results of this first preliminary classification indicated that about 60% of HERV could be placed either in class I (Gamma-like), class II (Beta-like) or class III (Spuma-like) and could be further assigned to, at least, 33 provisional clades (groups). However, it was not possible to properly execute a clear classification for the remaining 40% of the detected chains.

3.3. Evaluation of the unclassified HERVs

HERVs classification is largely based on polymerase (Pol) similarity (Andersson et al., 1999; Tristem, 2000) due to the highly conserved

reverse transcriptase and integrase portions and to the large size (around 1100 aa) of this protein. Thus, both Pol putative proteins (puteins, the amino acid sequences) and nucleotide sequences obtained from the 1308 still unclassified chains (using RvRef, RU and HML similarity) were used to reconstruct an unrooted Clustal guide tree. The backbone structure of this tree was also formed by the annotated set of HERV reference sequences (RvRef).

Next, using a tree-based clustering algorithm ("ConsensusFromTree", kindly performed by J. Blomberg) and the guide tree of the unclassified chains, 67 consensus sequences were generated. The algorithm allowed joining of chains into groups based on ocular inspection of the tree and calculation of consensus sequences from the group. The degree of heterogeneity of the group was also calculated. The procedure was performed both with trees based on Pol amino acid similarity and whole chain nucleotide sequence. These new consensus sequences were named according to the closest RvRef branch plus "x" (RvRefx).

Then, in order to complete the classification, a systematic BLAST analysis with the concatenated "Gag-Pro-Pol" amino acid sequences, as reconstructed by RetroTector, and the whole proviral nucleotide chains was also performed against the RvRef, the RvRefx and the "LTR" subset of the entire Repeatmasker (RM) collection of vertebrate repeats (release of May 2012) (Smit, 1996).

The results of this multi-step classification strategy allowed a better definition of the HERV clades. The retroviral chains could then be grouped in 56 clades (groups), the majority of them belonging to the Gamma-like (class I), in which a few Epsilon-like related elements were identified, Beta-like (class II) and Spuma-like (class III). The remaining clades were identified as MLT, MST and THE elements that are part of the large non-autonomous mammalian apparent LTR retrotransposon group (MaLR, class III), as well as a small number of non-LTR retrotransposons like LINE and SINE. Although most LINEs and SINEs were removed by "brooms", which were run by RetroTector prior to the main run, a small number of aberrant

representatives were still present after this sweeping procedure.

Applying this procedure, a majority of the chains could be unambiguously assigned to one specific clade, however, many proviral sequences showed a mosaic structure, derived from possible secondary integrations ("piggy-back") or other recombination events that complicate the HERV phylogeny. The handling of these ambiguities is difficult in term of phylogenetic reconstruction because the recombinant proviruses can confuse the analysis. To avoid these complications, the different components of each chain needed to be further evaluated.

3.4 Creation of Simage and analysis of mosaic elements

In order to resolve the complex genetic substructure of HERV sequences, a novel type of similarity image analysis (here named Simage) was performed (Fig. 3.1). As shown in Fig. 3.1, the retroviral chains (independently of their length) were sliced into twentieths and each of them was scored by BLASTn, using procedures kindly performed by J. Blomberg, against the entire HM, RvRef, RvRefx and RM references and consensus sequences collections, thus generating three different types of Simage per each chain: i) type 1 from the HML set; ii) type 2 from both the RvRef and rvRefx sets and iii) type 3 from the RM set.

A range of similarity between 0 and 9 was calculated and attributed to each twentieth. The retroviral chains for which more than ten twentieths derived from the same reference sequence and less than four twentieths were "0", were considered as unambiguous (or canonical) members of the most frequently matching reference sequence. In cases where both RvRef and RM indicated a different unambiguous reference sequence, preference was given to the RvRef sequence. The RvRef sequences can be traced to numerous HERV publications and are therefore important for maintenance of the collected knowledge on HERVs. However, the analysis with the RM system was performed simultaneously, so it was always possible to compare the two results.

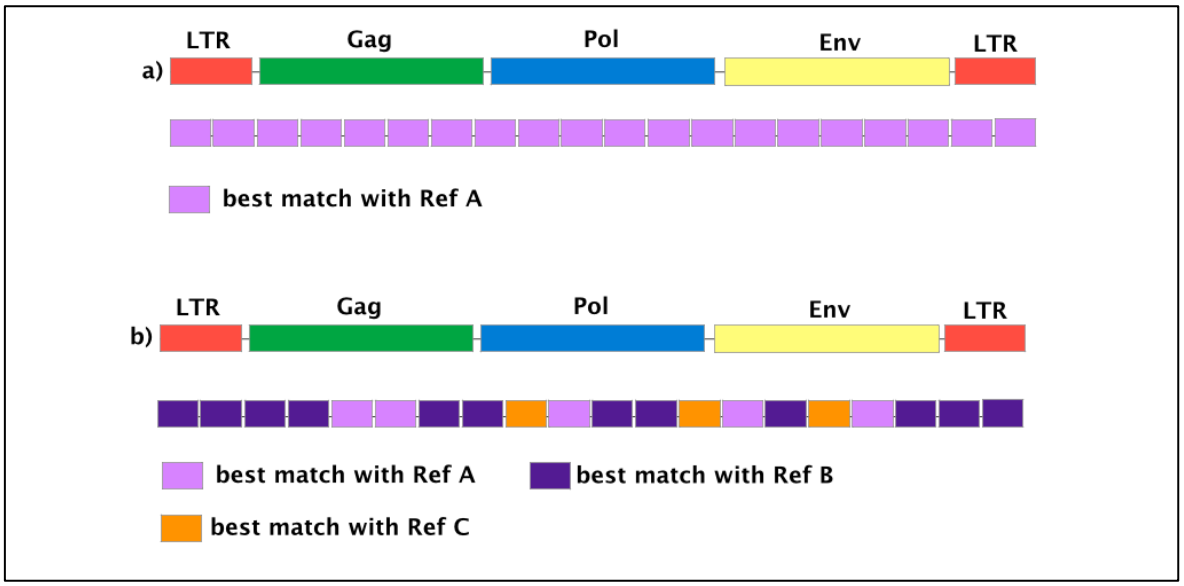

Fig. 3.1 Simage. Simplified representation of use of Simage in HERV classification. HERV chains, exemplified with a) and b), were sliced in twentieths and each twentieth was scored by BLASTn against different sets of retroviral reference sequences (here represented as Ref A, Ref B and Ref C). HERV chains were unambiguously classified if more than 50% of the twentieths matched the same reference sequence. In the example: sequence a) showed a 100% match with Ref A, whereas sequence b) had 60% of the twentieths matching with Ref B but short stretches of similarity with other two references (Ref A and Ref C) were also present.

Simage creation had a duplex value, on one hand the classification and the number of definitive HERV clades (groups) was reinforced with a clear definition of unambiguous canonical sequences and the level of homogeneity, on the other hand they showed that the presence of mosaic elements with heterogeneous content is higher than previously appreciated. Furthermore, the Simage analysis highlighted some issues concerning the generation and the handling of consensus sequences, such as the previous described RvRefx collection. In many cases, these consensus sequences turned out to be heterogeneous, probably due to recombination and/or secondary integration of LTR or non-LTR elements thus delimiting their application. A few RvRefx consensus sequences showed a more homogeneous profile, i.e. the "HEPSI" group, and could be then used for classification purposes.

The final results from the Simage (Fig. 3.2) analysis showed that among the 3290 identified HERV sequences, 1927 (about 60%) could be unambiguously assigned to a specific clade (canonical sequences) while

1363 (about 40%) could not be unequivocally classified to one clade (non-canonical sequences). However, these non-canonical sequences were provisionally assigned to the most frequent clade observed within the Simage. It is worth to note that the level of homogeneity inside the canonical chains in general is quite high. Most Simages concurred to a clear identification of the retroviral sequences.

Simages could be considered as a "magnifying glass" that allowed to look inside the proviral sequences. In this respect, it could also represent a preliminary tool for distinguishing the source of the heterogeneous content within the retroviral sequences.

Generally, Simage analysis led to the following important observations. Within the canonical classified sequences, a few chains (340) have short discrete stretches of similarity to Errantiviruses (gypsy elements) (Lynch & Tristem, 2003; Volff, 2009). In particular, this was observed inside the Pol portion (C-terminus) of the Gamma-like HERVW, HERVE, HERVH and HERVI clades that probably retained this portion because of an important function (Eickbush & Jamburuthugoda, 2008; Stefanov et al., 2012). The C terminal domain of the integrase is responsible for binding to chromatin.

An analogous observation could be done for other HERV sequences, but a one by one description was difficult to apply for the entire set of canonical sequences, given the large amounts of generated Simages. Therefore, in order to describe other unequivocal stretches of similarities, the creation of “consensus” Simage was then performed (Table 3.2).

The results showed that the most frequent stretches of similarities were recognizable within each type of Simage generated for the main canonical clades. Some of the observed stretches could be explained with the phylogenetic relationships and the evolution of these clades, for example it was not surprising to recognize MER41 and ERV9 stretches of similarity in some sequences within the HERVW clade or stretches most similar to HML9 within the HML1 clade.

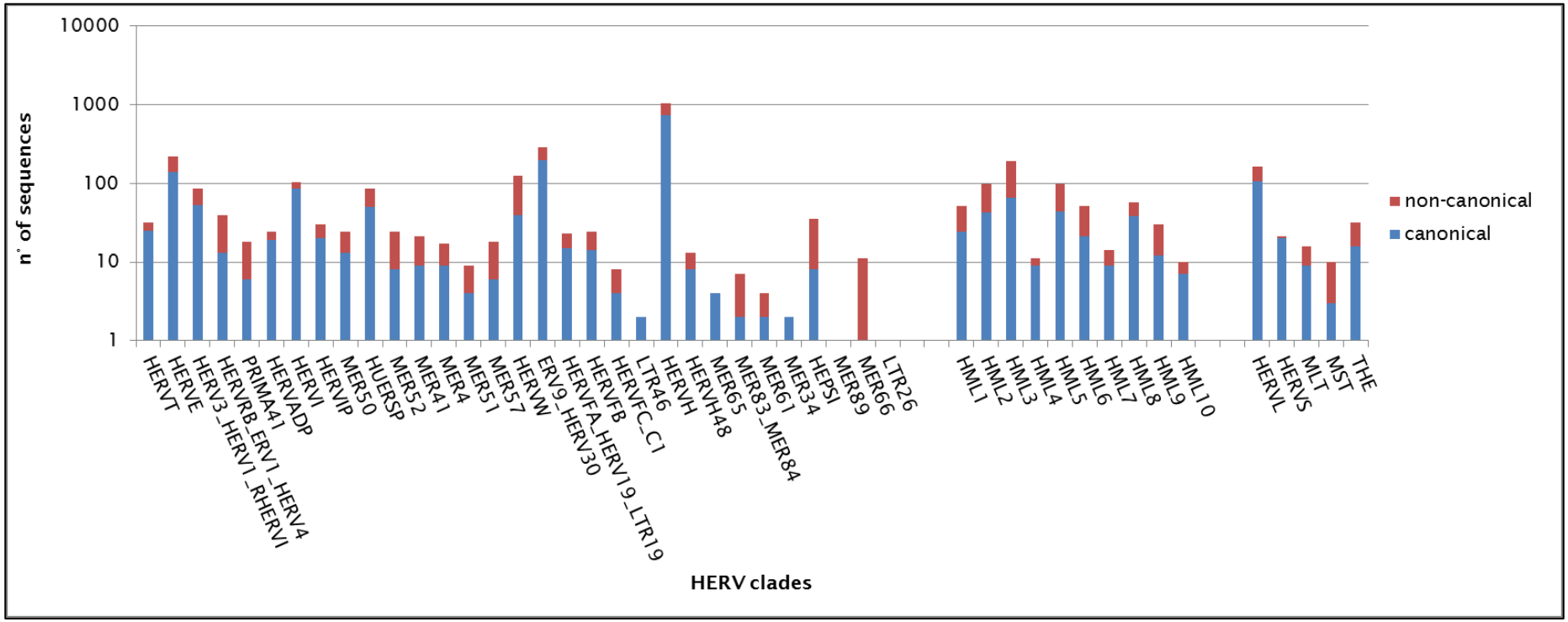

Fig. 3.2 Classification of HERV sequences identified in GRCh37/hg19. According to Simage classification procedure, the HERV sequences were distinguished in canonical (blue bars) and non-canonical (red bars).

Table 3.2. Description of most frequent rearrangements within some canonical HERV clades

HERV clades	N° of detected sequences	§Simage types		
		type 1	type 2	type 3
HML1	24	HML6/HML9 (x8)	ND[b]	HML9 (x6)
HML2	43	HML10/HML1 (x7)	ND[b]	HML3/HML7/HML9 (x17)
HML3	65	HML1 (x8); HML2 (x5)	HML4 (x61)	HML9 (x7)
HML5	44	HML10 (x3)	HML10 (x7)	HERV9 (x3)
HML8	38	ND[a]	HML1(x3); HML4 (x2)	ND[c]
HERVH	744	NA[d]	HUERSP (x142)	MST/MERX[e] (x22); MLT (x40)
HERVW	39	NA[d]	MER41 (x13); ERV9 (x26)	ERV9 (x8)
HERVI	105	NA[d]	HERVE (x90)	HERVW (x9)
HERVE	139	NA[d]	HERVT (x17); PRBX[e] (x10)	ND[c]
ERV9	188	NA[d]	HERVW/HMLX[e] (x94)	HERVW/ HERVIP (x22)

§Simage types are so defined: type 1 is based on HML set of consensus sequences (generated by Blikstad V.); type 2 is based on both the RvRef and RvRefx sets of retroviral references (from literature) and consensus sequences (generated in this study); type 3 is based on LTR RM (Repeatmasker) collection. Frequencies of most observed stretches in a clade are given in parentheses. Simultaneous stretches are indicated by "/", independent stretches are separated by ";" [a]Simages generated using the HML set of consensus sequences were not homogeneous to define the HERV clades; [b]Simages generated with the RvRef and RvRefx sets of reference and consensus retroviral sequences were not homogeneous to define HERV clades; [c]Simages generated with the RM set of consensus sequences were not homogeneous to define HERV clade; [d]HML Simages were not applicable to Gamma-like HERV clades; [e]different types of MER or PRB or HML are included.

However, some Simages turned out to be more heterogeneous than others, thus limiting the generation of a "consensus" Simage for a number of HERV clades. This can be explained by an intrinsic property of Simages. They are based on reference and different consensus sequences. Generally, each consensus sequence, occurring in the RM, HML and the RvRefx collections, represents an "average" of several more or less complete sets of retroviral sequences. The heterogeneity of a set represented by a single consensus can give rise to the observed variability within the respective Simages. The RvRefx collection was only based on unclassified chains from the GRCh37/hg19 assembly of the human

genome. In contrast, the RM consensus collection covered a wide panel of species-specific variants of retroviral sequences from different vertebrates. This naturally lead to an apparent greater heterogeneity of the RM Simages where closely related ERVs from different species sometimes occurred in a Simage, erroneously indicating a greater heterogeneity than they effectively had.

Nonetheless, the consensus Simages showed that, within the main HERV clades, the level of rearrangements is quite low and is confined to highly similar members (beta/beta, gamma/gamma) thus confirming the level of unambiguous classification of the sequences.

Simage analysis did however reveal mosaic sequences, sometimes of uncertain origin. In some cases this was traceable to secondary integrations into a proviral sequences, while in other cases retroviral recombination before integration, during reverse transcription, could be responsible (Stuhlmannt & Berg, 1992) (Hu & Temin, 1990). The proviral sequences that were not unequivocally assigned to a specific clade but with a backbone structure roughly identified as Gamma-, Beta- or Spuma-like were evaluated for the presence of different HERV components. A preliminary analysis of these Simages showed that a wide fraction of the mosaic sequences harbor a high presence of MaLR (MST, MLT and THE) within a gamma or beta HERV backbone. Secondary integrations of HML2 or HML8 LTRs occurred both in Gamma- and Beta-like chains belonging to other clades. A further evaluation of these secondary integrations could be of particular interest since the fact that one of the major roles of HERV sequences in shaping the host genome functions is linked to the presence and position of single-LTR that could account for alternative promoters and/or enhancers (Belshaw et al., 2007; Cohen et al., 2009).

As mentioned, sequence heterogeneity could arise either from secondary integrations, recombination events or from similarity to several evolutionary related and conserved retroviral sequences (i.e. the Pol gene). In order to trace the origins of this heterogeneity, a bioinformatics recombination analysis with the SimPlot software (Lole & Ray, 1999) is

currently under development.

Simages allowed a phylogenetic reconstruction of homogeneous HERV sequences, avoiding misclassification due to sequence mosaics.

3.5 Phylogenetic trees

The best representative Pol and Gag putative proteins (puteins) generated by RetroTector from the classified HERVs were selected to generate neighbor joining (NJ) guide trees (Fig. 3.3 and 3.4). A broad panel of retroviral reference sequences was also included in both trees, for taxonomic purposes.

The phylogenetic trees essentially confirmed the classification shown in Fig. 3.2, with the clusterization of HERV sequences in the three canonical and well-described retroviral genera (Gamma-like, Beta-like and Spuma-like). Gamma-like can be further delineated in two subgroups, based on the presence of two or one zinc-finger motifs in their nucleocapsid (NC) portion of Gag protein, respectively. The presence of one zinc-finger motif is a signature of both exogenous and endogenous Gammaretroviruses (MLV, HERVE, HERVW). However, a second zinc-finger, more or less complete, was described within the HERVH group, thus leading to the hypothesis of an evolutionary divergence with the differentiation in "old" (two zinc-finger) and "young" (one zinc-finger) Gammaretrovirus groups (Jern & Sperber, 2005a).

Moreover, some adjustments could be done within the Gamma-like HERV clades. In some cases, HERV clades previously classified by Simages could be further merged with the closest clade branch (i.e. ERV1_cow with HERVRB, HERV30 with ERV9).

In addition to the canonical retroviral genera, the Pol and Gag phylogenetic trees highlighted the presence of a small group of Epsilon-like sequences close to the most basal branches of the Class I Gamma-like group. The classification of these sequences, as emerged from the Simage as well as

from the phylogenetic reconstruction, seems to confirm their relationships with the exogenous member of the Epsilonretrovirus genus, WDSV, and justifies the classification of these sequences as a HERV clade (here named “HEPSI”) on its own (Jern et al., 2005b; Oja et al., 2005).

Another interesting observation can be done for the HERVL members that are considered as class III HERV, distantly related to the Spuma retroviruses (Cordonnier et al., 1995). In the Pol tree, one HERVL Pol putein, at least, clusters with the Epsilon-like elements thus confirming the previous described presence of an intermediate group between Gamma-like and Spuma-like elements (Jern et al., 2005b) and the more complex role of MuERVL as progenitor of Epsilon-like particles (Ribet et al., 2008).

Finally, the defined HERV clades were analyzed for their degree of heterogeneity. This issue was achieved through the generation of consensus sequences by joining the whole nucleotide chains and the most intact puteins (Gag, Pro, Pol and Env) within each of the classified HERV clade. In order to achieve a 50% of both "intermember identity" (identity within the group, IWIG) and "identity to consensus" (ITC) within the groups, each set of HERV chains within the clade were joined to generate the definitive consensus sequences, thus condensing HERVs into a small sequence set which useful for classification purposes and identification of unknown sequences occurring in large scale sequencing efforts.

This phylogenetic analysis enabled refinement of clade assignments, leading to the final HERV classification.

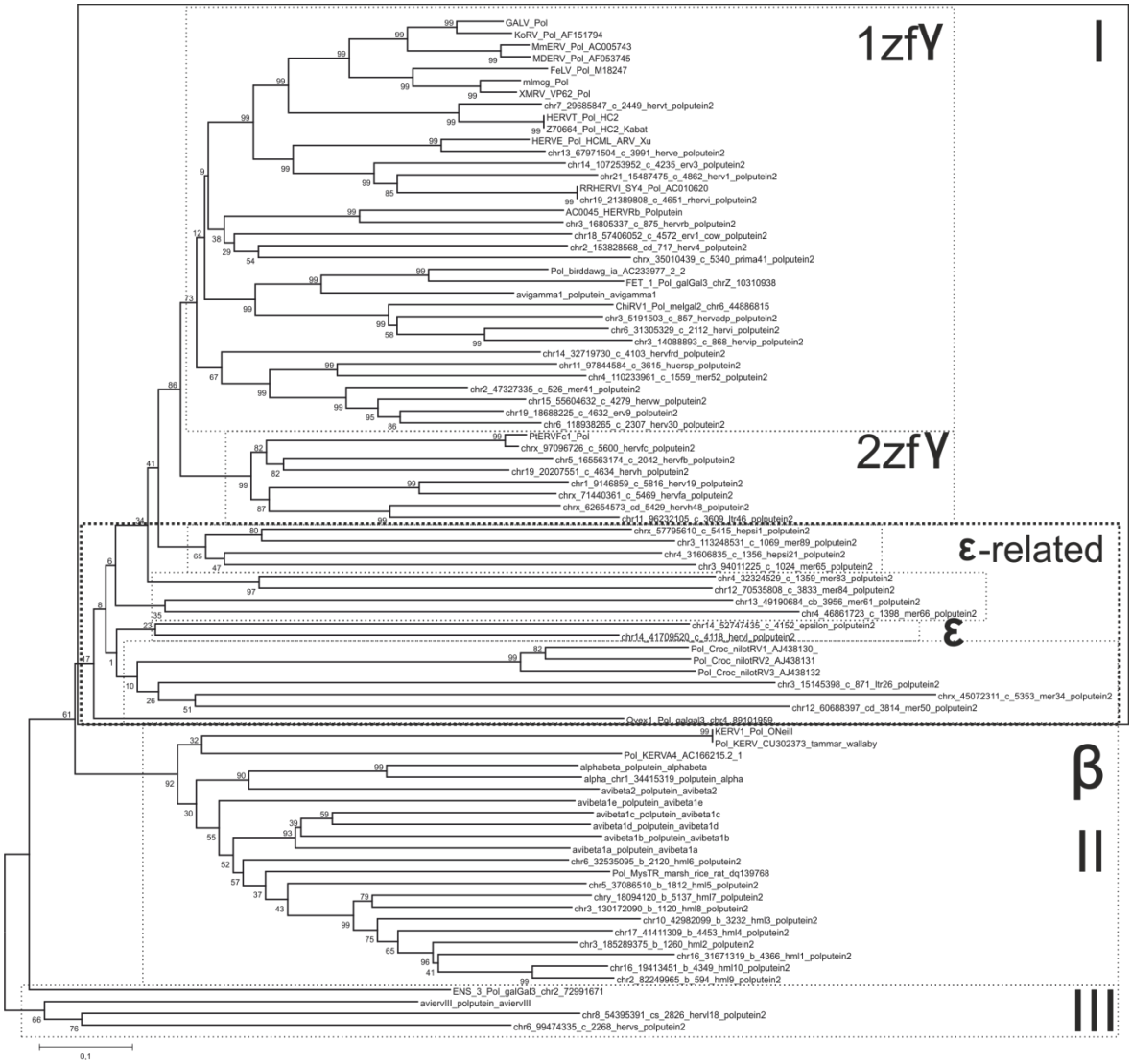

Fig. 3.3 Phylogenetic analysis of HERV proviruses. Unrooted neighbor joining (NJ) tree of selected Pol proteins (puteins) from HERVs found in GRCh37/hg19

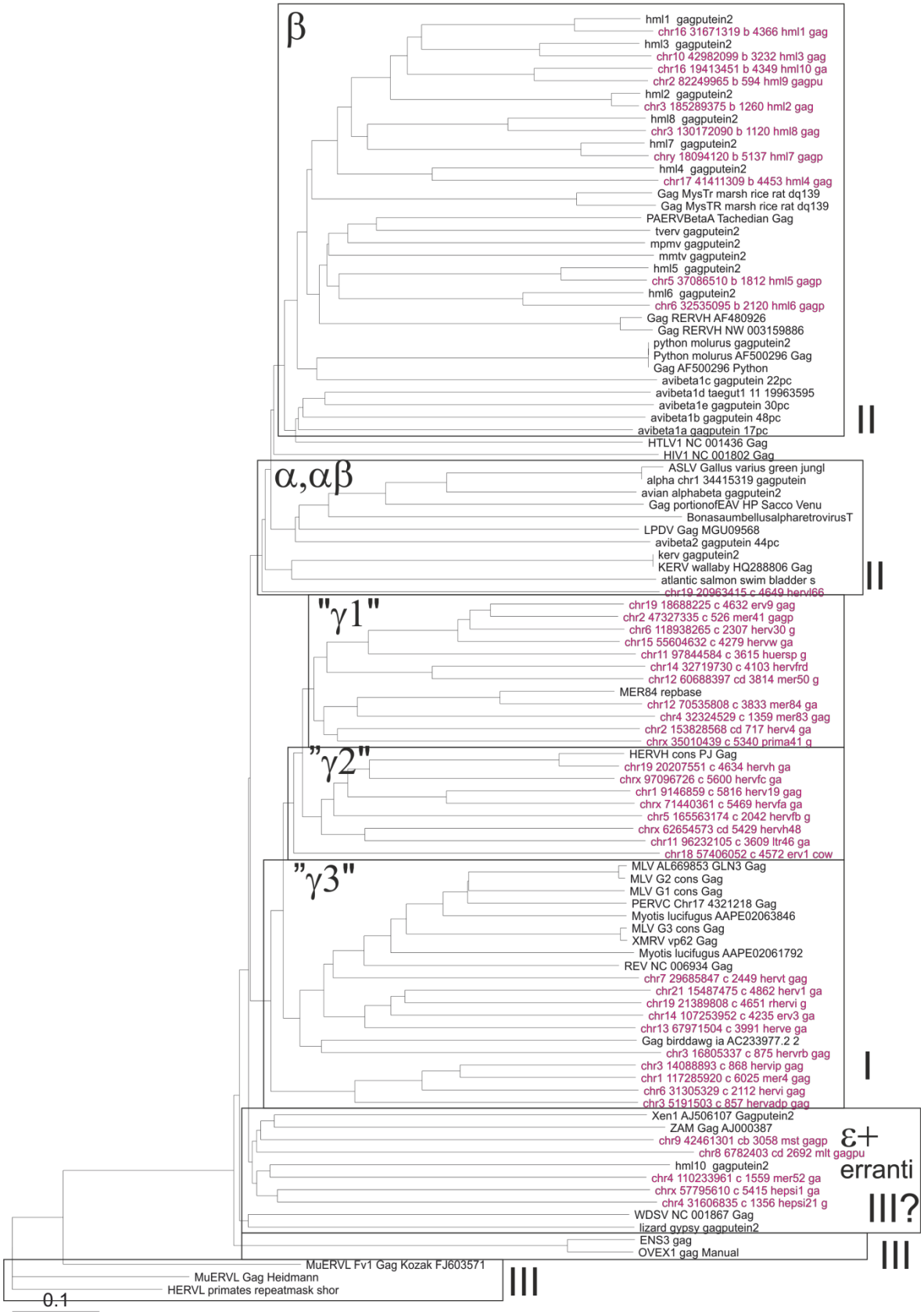

Fig. 3.4 Phylogenetic analysis of HERV proviruses. Unrooted neighbor joining (NJ) tree of selected Gag proteins from HERVs found in GRCh37/hg19. Magenta color indicates HERVs from this study; black color indicates exogenous and endogenous reference sequences (RvRef).

3.6 Final HERV classification emerging from the multistep procedure

The final number of proviral sequences and the definitive HERV clades (groups) assignments, as emerged from the above described multistep classification procedure is described in Table 3.3 and 3.4. About 95% of the 3290 proviral sequences identified in GRCh37/hg19 can be assigned into the class I (Gamma- and Epsilon-like), II (Beta-like) and III (Spuma-like, MaLR group) whereas most of the remaining few elements, that were not clearly class I-III HERVs, were classified as non-LTR retrotransposons (Table 3.3).

A total of 40 HERV clades are listed in Table 3.4, in which only the number of the canonical classified elements per each clade is listed. Table 3.4 also shows the most common HERV groups and the estimated copies number previously reported in literature (Mager & Medstrand, 2003). The strict comparison between the two sets of data, both in terms of nomenclature and copies number, is not easy to achieve because of the different strategies occurred in the HERV identification and classification. However, some data are confirmed and can be compared mainly for those clades that are well-characterized (i.e. HERVW, HERVH or HERVK(HMLx), whereas the data defined for a large number of clades are more precise than previously reported.

Table 3.3. Final classification of HERVs identified in GRCh37/hg19

[a]Genera	Species	HERV genus	[b]N° of total sequences	[c]N° of clades
Gammaretrovirus	Murine leukemia virus (MLV) Moloney murine sarcoma virus (MoMuLV) Feline leukemia virus (FeLV)	Gamma-like	2298	24
Epsilonretroviruses	Walleye dermal sarcoma virus (WDSV)	Epsilon-like	35	1
Betaretrovirus	Mouse mammary tumor virus (MMTV) Mason-Pfizer monkey virus (MPMV) Jaagsiekte sheep retrovirus (JSRV)	Beta-like	615	10
Spumaretrovirus	Simian foamy virus (SFV)	Spuma-like	183	2
		[d]MaLR (i.e. MST-MLT-THE)	58	3
		[e]Others	94	-
		Tot	**3290**	**40**

[a]Retroviruses classification from International Committee on Taxonomy of Viruses (ICTV) 2012; [b]number of HERV sequences identified by RetroTector, values are comprehensive of both canonical and non-canonical classified HERV sequences (see details in the main text); [c]number of final classified HERV clades (groups); [d]according to the Repbase classification system, the MaLR elements (MST; MLT and THE) belong to class III ERVs; [e]LINE, SINE and other MER sequences not proper belonging to class I-III HERV (see details in the main text).

Table 3.4 List of 40 canonical HERV clades found in GRCh37/hg19

HERV clade	[1]N of identified HERV sequences	[2]Common name (Repbase identifiers)	[3]Estimated copies
Class I			
HERVT	25	HERVT (HERVS71/LTR6)	80
HERVE	139	HERVE (HERVE/LTR2)	250
HERV3	53	HERV3 (HERV3/LTR4)	100
		RHERVI (HERV15I/LTR15)	40
HERVRB	13	HERV-Rb (PABL_BI/PABL_A, PABL_B)	8
PRIMA41	6	PRIMA 41(PRIMA41/MER41)	40
HERVADP	19	HERVADP (HERVP71A_1/LTR71)	40
HERVI	105	HERVI (HERVI/LTR10)	250
MER50	13	HERVFRD (MER50I/MER50)	50
HUERSP	50	HERVP (HUERSP3/LTR9)	200
MER52	8	MER52A (MER52AI/MER52A)	200
MER4	9	(MER4)	NA
MER51	4	(MER51)	NA
MER57	6	HERVHS49C23 (MER57I/MER57)	200
HERVW	39	HERVW (HERV17/LTR17)	40
HERV9	204	ERV9 (HERV9/LTR12)	300
HERVFA	15	HERVF (HERVFH19/LTR19)	45
HERVFB	14	HERVFXA (HERVFH21/LTR21A)	30
HERVFC	4	HERVFC (HERV46I/LTR46)	6 (Bénit et al., 2003)
HERVH	736	HERVH (HERVH/LTR7)	1000
HERVH48	8	HERVFb (HERVH48I/MER48)	60

[1]Number of canonical classified HERV sequences identified in this study (see details in the main text); [2]see: Bannert & Kurth, 2006 and Mager & Medstrand, 2003; [3]if not differently stated, see: Mager & Medstrand, 2003.

Table 3.4 List of 40 canonical HERV clades found in GRCh37/hg19 (continued)

HERV clade	[1]N of identified HERV sequences	[2]Common name (Repbase identifiers)	[3]Estimated copies
MER65	4	(MER65)	NA
MER84	2	MER84 (MER84/MER84)	25
MER61	2	(MER61)	NA
MER34	2	(MER34)	NA
HEPSI	8	NA	NA
Class II			
HML1	24	HML1 (HERVK14I/LTR14)	70
HML2	43	HERVK(HML2) (HERVK/LTR5)	91 (Subramanian et al., 2011)
HML3	65	HML3 (HERVK9I/MER9)	150
HML4	9	HML4 (HERVK13I/LTR13)	10
HML5	44	HML5 (HERVK22/LTR22)	100
HML6	21	HERVK(HML6) (HERVK31/LTR3)	50
HML7	9	HML7 (HERVK11DI/MER11D)	20
HML8	38	HML8 (HERVK11I/MER11A)	60
HML9	12	HERVK(14C)/NMW9	10
HML10	7	HML10 (HERVKC4/LTR14)	10
Class III Spuma-like			
HERVL	105	HERVL (HERVL/MLT2)	200
HERVS	20	HERVS (HERV18/LTR18)	50

[1]Number of canonical classified HERV sequences identified in this study (see details in the main text); [2]see: Bannert & Kurth, 2006 and Mager & Medstrand, 2003; [3]if not differently stated, see: Mager & Medstrand, 2003.

Table 3.4 List of 40 canonical HERV clades found in GRCh37/hg19 (continued)

HERV clade	[1]N of identified HERV sequences	[2]Common name (Repbase identifiers)	[3]Estimated copies
Class III MaLR group			
MLT	9	MLT	NA
MST	3	MST	NA
THE	15	THE	NA

[1]Number of canonical classified HERV sequences identified in this study (see details in the main text); [2]see: Bannert & Kurth, 2006 and Mager & Medstrand, 2003; [3]if not differently stated, see: Mager & Medstrand, 2003.

3.7 Discussion

In spite of the great efforts made during the last 25/30 years, a comprehensive analysis of the most intact HERV proviruses present in the human genome was still missing. Moreover, the main HERV databases (Paces et al., 2004; Villesen, Aagaard, Wiuf, & Pedersen, 2004) are not still maintained and updated. These issues could complicate the findings of basic information regarding HERV features, such as the number, the chromosomal position and an (more or less) in depth description of the genetic structures of the proviral sequences.

The main objective of this project was the characterization of HERV proviruses found in the GRCh37/hg19, and it represents an important step forward in the state-of the-art HERV research. In order to identify and characterize the retroviral sequences integrated in the human genome, the bioinformatics approach with the RetroTector software was used. This allowed us to avoid any misleading and/or missing information on the endogenous retroviral sequences publicly available but difficult to retrieve.

RetroTector allowed the identification of 3290 HERV actually integrated in the human genome, one of the latest and most thoroughly made assemblies. The dataset included not only the complete nucleotide sequences but also a panel of parameters, such as the nucleotide length, the chromosomal position (start and end of the sequences), the presence of, more or less complete, sub-motif hits such as retroviral genes (Gag-Pro-Pol-Env) or other structural features, i.e. primer binding site (PBS) or polypurine tract (PPT).

The multistep classification procedure led to the taxonomic identification of about 95% of the 3290 HERV with the expected over-representation of Gamma-like sequences with respect to the Beta-like and the total absence of recognizable Alpha-, Delta- or Lentiviral-like proviral sequences. However, the presence of few Epsilon-like elements is worth of note and will deserve a more detailed investigation.

In general, the 40 HERV clades identified are comparable with those previously described and the description of the most integer, and

undoubtedly classified, proviral elements is surely reinforced by the development of the Simage analysis. Simages defined the complex genetic structure of the 3290 HERV with a level of detail not previously appreciated. HERV are generally described as fragmented, deteriorated remnants of their exogenous retroviral ancestors but the exact contributions, in terms of nucleotide sequences, to the mosaic structure of most of the HERV is here firstly described, but needs further study.

4. Analysis of identified HERVs

4.1 Introduction

The identification and classification of HERV sequences, the "Who" described in Chapter 3, was a preliminary and necessary step before proceeding to a most extensive characterization and a general description of the context of the proviral integrations within the human genome GRCh37/hg19, the "Where". Indeed, the roles of HERV sequences in shaping and regulating the human genome and, moreover the link between HERV and diseases, are still vague and need to be further investigated (Kurth & Bannert, 2010; Magiorkinis et al., 2013).

The analysis of the HERV sequences, described in this chapter, is divided in three main parts. The first one (subchapters 4.2 and 4.3) is about the investigation of a wide panel of conserved and recognizable retroviral genetic traits such as, e.g., the number of Gag nucleocapsid (NC) zinc-fingers, the translational strategy and the primer binding site (PBS) usage. These structural markers were analyzed with the aim to confirm both the congruency and the homogeneity of the previously performed HERV clades classification. In fact, it is known that many Gammaretroviruses share the presence, in the nucleocapsid portion of the Gag protein, of a one zinc-finger motif whereas Betaretroviruses have two zinc-finger motifs and Spumaretroviruses none (Freed, 2002; Mirambeau et al., 2010). In a previous report, a similar pattern was also observed in the endogenous retroviruses where Gamma-like sequences could be further ascribed to "one" (i.e. HERVT, HERVW and HERVE) and "two zinc-finger" groups (HERVH and HERVH related) (Jern & Sperber, 2005a; Jern et al., 2005b), respectively.

Similarly, the putative HERV translational strategy was also reconstructed keeping in mind that Retroviruses are known to regulate their protein expression either *via* a read-through suppression of the termination codon located at the end of the Gag frame (Gamma and Epsilonretroviruses) or by ribosomal frameshifting strategy (Betaretroviruses). In both cases, these

different translational strategies lead to the synthesis of a Gag-Pro-Pol precursor (Coffin, 1997). In contrast, in Spumaretroviruses the Pro and Pol proteins are expressed independently from Gag, thus implying a different use of frameshifts between the Gag-Pro and Pro-Pol boundaries (Enssle et al., 1996; Lee et al., 2013).

Finally, a further characterization of the 3290 classified HERV sequences could also be achieved through the analysis of their PBS sequences complementary to the tRNA used to prime the viral genome reverse transcription. Indeed, HERV groups were historically classified after their PBS, i.e. HERVW was named after recognition of a PBS sequence sharing homology with the avian retrovirus $tRNA^{Trp}$ (Blond et al., 1999). However, in some cases this correlation could not reflect the real PBS usage for all the members of the same clade (Blomberg et al., 2009; Jern et al., 2004), thus a finest investigation of the HERV PBS sequences, identified by RetroTector, was performed.

The second part of the chapter (subchapter 4.4) is focused on the analysis of the protein coding potential retained by the 3290 identified HERV sequences. Addressing this issue could be of particular interest to establish the basis of a transcriptome database/collection of putative active retroviral proteins (Flockerzi et al., 2008). Hence, a description of the total number of ORFs for all the retroviral genes (Gag, Pro, Pol and Env), as reconstructed by RetroTector, together with statistical data analyses of their completeness (number of stop codons and frame-shifts) was elaborated.

The last part of the chapter (subchapter 4.5 and 4.6) describes the overall investigation of the background of HERV sequences integrations within the human genome GRCh37/hg19 with a particular focus on their chromosomal distribution (clusters analysis) and proximity to transcription units (TUx).

4.2 Characterization of HERV structural markers

The HERV sequences classified as canonical were analyzed with respect to a series of motif hits with the aim to complete, as far as possible, the genetic structure of the retroviral chains.

The HERV sequences were searched for the follows structural markers: i) NC zinc-finger motifs, ii) translational strategy, iii) $dUTP^{Pro}$ and $dUTP^{Pol}$, iv) C-terminal motifs of protease and polymerase genes (G-patch and GPY/F) and v) nucleotide biases. The results of this broad survey are summarized in Table 4.1 and Table 4.2.

The data shown in Table 4.1 complied, to a large extent, with those reported in literature both for exogenous and endogenous retroviruses. In fact, no main outlier within the HERV genus could be detected in $dUTP^{Pro}$, G-patch, GPY/F and nucleotide values, that are some of the known structural markers for Beta-like and Gamma-like retroviruses, respectively (Jern et al., 2005b; Mayer & Meese, 2003). However, it is worth to note that the $dUTPase^{Pol}$, a signature previously reported for HERVL and MuERVL (Cordonnier et al., 1995; Bénit et al., 1997) was not detected by RetroTector in the 1.01 utilized version. Instead, the presence of this genetic trait within the classified HERVL sequences was confirmed by the inspection of the Simages (see Chapter 3.4) derived from their Pol proteins. As reported in Table 4.1, the results confirmed the presence of a $dUTPase^{Pol}$ within all the members of the HERVL clade, thus supporting the previous classification. It is also worth to highlight the total absence of the same marker within the other major Spuma-like group, the HERVS clade.

Table 4.1 Summary of HERV structural markers in human genome GRCh37/hg19

HERV genus	[a]N°	Pro	Pol	C-terminal Pro	C-terminal Pol motifs		Nucleotide biases (>25%<)	
		dUTPase	dUTPase	G-patch	IN detected region	GPY/F		
Gamma-like	1477	0	1 (HERVH)	1 (HUERSP)	677	369	↑AT ↓G	[b]↑C ↓G HERVH and HERVH-related
Epsilon-like	8	0	0	0	8	6	↑AT	↓G
Beta-like	272	195	3 (HML2)	65	191	8 (HML6)	↑A	NA
Spuma-like	128	0	[b]100 (HERVL)	[b]2 (HERVL)	59 (HERVL ; HERVS	29 (HERVL)	NA	NA

[a]Number of canonical classified HERV sequences (see Chapter 3); [b]outlier (or new) results for particular HERV clades are shown.

As far as the presence and the number of zinc-finger motifs are concerned, the results are as expected, but new data arose from the survey reported in Table 4.2. A total number of 671 out of the 736 HERVH classified sequences had a reconstructed Gag with a NC portion. Among them, 528 HERVH elements were found to harbor one zinc-finger motif instead of the previously described two zinc-fingers motifs. Indeed, only 200 HERVH sequences were found to follow the two zinc-finger rule, a pattern that was observed also in other HERVH related groups, such as HERVH48 and HERVF sequences that overall shown a strong two zinc-finger bias. The further alignment of the 200 HERVH Gag proteins together with a panel of Gag reference sequences, showed that 144 out of 190 had a second zinc-finger most similar to that of a Deltaretroviral zinc-finger whereas 53 out of 200 had a second zinc-finger most similar to a Betaretroviral one.

A second result worth to be highlighted is the presence of other Gamma-like sequences that showed a two zinc-finger trend, a data not previously investigated. This is the case of 17 ERV9 sequences that had a

recognizable second zinc-finger that shares sequence similarity mainly with the Betaretroviral zinc-finger motif.

Table 4.2 Summary of HERV structural markers in human genome GRCh37/hg19

HERV genus	[a]N°	NC zinc-fingers		[b]Gag-Pro f.s.			[b]Pro-Pol f.s.			[b]Shift-strategy
		1	2	-1	0	+1	-1	0	+1	
Gamma-like	1477	1038 [c](528 in HERVH)	266 [c](17 ERV9)	88	**154**	93	241	**313**	189	0/0 [c]0/+1 in HERVI [c]+1/-1 in HERVW
Epsilon-like	8	7	0	4	0	0	0	0	0	[d]ND
Beta-like	272	52	161	**32**	16	12	**45**	22	17	-1/-1
Spuma-like	128	33 [c](25 in HERVL)	1	21	10	16	26	33	21	[c]-1/0 in HERVL
Tot	1885	1130	428							

[a]Number of canonical classified HERV sequences (see Chapter 3); [b]predicted frameshift (f.s.) translational strategy between the putein ORFs boundaries (Gag-Pro and Pro-Pol); [c]outlier (or new) results from specified HERV clades are shown; [d](ND): not determined. The few Epsilon-like detected sequences showed a high rate of stop codons and frameshifs in the main ORFs, thus avoiding to perform the analysis.

Another surprising data is about the Spuma-like (HERV class III) retroviral chains. Some of the classified HERVL and HERVS sequences were found to have one NC zinc-finger, whereas it is assumed that the elements of this retroviral genus do not harbor any zinc-finger motif (Bowzard et al., 1998; Flugel, 1991). This should be further investigated.

The results of HERV translational strategy are also reported in Table 4.2. The retroviral proteins (puteins) from the most intact and canonically classified HERV sequences were analyzed for their reading frames and the translational strategies were calculated. In general, the Gamma-like (HERV class I) retroviruses are devoid of frameshifts, "0 f.s.", both in Gag-Pro and Pro-Pol boundaries, thus confirming the tendency for these proviruses to translate their proteins in the same ("0/0") reading frame. However, within the Gamma-like sequences a few elements differed from the others, i.e.

HERVW and HERVI clades that showed a ("+1/-1") and a ("0/+1") frameshift pattern, respectively. These "deviations" from the expected Gamma-like strategy should be further investigated. Defective proviral sequences with postintegrational deletions and/or shifts inside genes could have contributed to the observed unexpected proviral reading-frames.

Within the Beta-like HERV (HERV class II) analyzed, the results complied with those observed in the exogenous counterparts (i.e. MMTV) which are known to prefer the ribosomal frameshifting strategy. Indeed, a "-1/-1 f.s." could be observed between Gag-Pro and Pro-Pol reading-frames.

As mentioned above, the Spumaretroviruses, unlike the other retroviruses, produce a Pro-Pol fusion protein from a spliced mRNA, however, less is known about the more distantly related endogenous Spuma-like retroviruses. As reported in Table 4.2, an attempt to reconstruct the translational strategy of these sequences was also performed. This was possible with respect to the only HERVL clade, which showed a "-1/0 f.s." translational strategy. Thus, these sequences seem to resemble the related Spumaretroviruses in term of hypothetical protein expression control.

4.3 HERV PBS sequence analysis

RetroTector identified a total number of 2200 PBS sequences within the whole set of 3290 classified HERV showing a PBS score range of 0-200. According to a preliminar (first) RetroTector classification, the PBS could be ascribed to the following main types (identified by a single-letter amino acid code): His (H=746), Arg (R=215), Pro (P=206), Phe (F=178) and Lys (K=167). However, a quality control of the RetroTector procedure of PBS sequence detection was needed to avoid misleading information.

The whole PBS sequence set, detected by RetroTector, was BLAST scored against the *H. sapiens* subset of the entire Leipzig tRNAdb collection (Juhling et al., 2009). The results are summarized in Table 4.3. Some interesting observations could be done: i) the most frequent start codon for the PBS sequences resulted to be "TGG", whereas "TTG" was mainly

recovered within the HERW and ERV9 clades; ii) the HERVL clade to which was previously assigned, by RetroTector, a Met (M) PBS type turned out to have a Leu (L) PBS type (as expected) and iii) the PBS usage could be rather different from the data generally reported in the literature. The latter case can be exemplified by ERV9 and HERVW that showed PBS sequence homologies either to an Arg (R) or a Trp (W) tRNA, in both clades. Another outlier is ERV3 that seems to use a Pro (P) PBS more frequently than an Arg (R) PBS, as normally reported (in fact, it was also named HERV-R) (Andersson et al., 2005).

Surely, the most outstanding data is the presence of a subset of HERVH that seems to use a Lys (K) PBS instead of the normal His (H). Phe (F) PBSs were also relatively frequent (Jern et al., 2004). The BLAST analysis of these Lys PBS sequences against the *H. sapiens* subset of tRNA allowed the distinction of two sets: 1) the "TGGTCTCTTCACATGGAC" ($tRNA^{Lys}$, with codon TTT), identified by RetroTector as a low score Lys PBS in 11 HERVH elements and 2) the "TGGTCTCTTCACACGGAC" ($tRNA^{Lys}$, codon TTT) identified as a low score Trp (W) PBS in 8 HERVH retroviral chains. It is noteworthy that the above-described Lys PBSs differ from those used by the HML2 clade. In effect, HML2 sequences showed both a "TGGTGCCCAACGTGGAGG" (tRNALys, codon CTT) and a "TGGCGCCCAACGTGGGGC" ($tRNA^{Lys}$, codon CTT). The different PBS usages among the HERVH sequences are also illustrated in Fig. 4.1, which shows a neighbor-joining guide tree with the relationships between the HERVH sequences based on their PBS usage.

Table 4.3 Selected PBS sequences within HERV clades

HERV clade	[a]PBS			[b]BestPBSReTe		[c]BesPBStRNAdb	
	sequence	[d]nr	[e]type	sequence	[e]type	sequence	[e]type
HERVH	[f]**TGG**TGCCGTGACTCGGAT	203	H	TGGTGCTGTGACTCAGAT (16)	H	TGGTGCCGTGACTCGGAT (18)	H
HERVH	TGGTGCTGAAGACCCGGG	7	F	TGGTGCTGAGACCCGGGA (13)	F	TGGTGCCGAAACCCGGGA (13)	F
HERVH	TGGTCTCTTCACACGGAC	9	W	TGGCGCCCGAACAGGGAC (11)	K	TGGTGCCTGAACAGGGAC (11)	K
HERVH	TGGTCTCTTCACATGGAC	11	K	TGGCGCCCGAACAGGGAC (11)	K	TGGTGCCTGAACAGGGAC (13)	K
HML2	TGGCGCCCAACGTGGGGC	8	K	TGGCGCCCAACGTGGGGC (18)	K	TGGCGCCCAACGTGGGGC (18)	K
HML2	TGGTGCCCAACGTGGAGG	16	K	TGGCGCCCAACGTGGGGC (15)	K	TGGCGCCCAACGTGGGGC (15)	K
HERVW ERV9	TGGCAACCACGAAGGGAC	12	W	TGGCAACCACGAACGGAC (17)	W	TGGTGACCCCGACGTGAC (13)	W
ERV9 HERVW	[f]**TTG**GTGACCACAAAGGGA	9	R	TTGGCGACCACGAAGGGA (16)	R	TGGTGACCCCGACGTGAT (13)	W
ERV9 HERVW	TTGGTGACCACGAAGGGA	9	R	TTGGCGACCACGAAGGGA (17)	R	TGGTGACCCCGACGTGAT (14)	W
ERV9 HERVW	TTGGTGACCATGAAGGGA	7	R	TTGGCGACCACGAAGGGA (16)	R	TGGTGACCCCGACGTGAT (13)	W
MER41 ERV9	TTGGCGACCCAGATGGGA	9	R	TTGGCGACCACGAAGGGA (15)	R	TGGCGAGCCAGCCAGGAG (13)	R
HERVL	TGGTACCAGGAGTGGTTC	15	M	TGGTGTCAGAAGTGGGAT (12)	L	TGGTGTCAGGAGTGGGAT (13)	L

[a]Final PBS sequences identified by RetroTector after the 3'-end of the 5'-LTR retroviral sequences; [b]best match of the final PBS sequence against the RetroTector PBS collection, in parenthasis the number of identical nt positions; [c]best match of the final PBS sequence against the tRNAdb collection, in parenthasis the number of identical nt positions; [d]number of times that the given PBS sequence was detected within the HERV clades; [e]single letter aminoacid code identifier; [f]in bold the most common codon start "TGG" and "TTG" identified among the final PBS sequences.

Fig. 4.1 Neighbor-joining tree of representatives HERVH PBSs. Red color indicates HERVH sequences with K TTT PBS (tggtctcttcacatggac); Magenta indicates HERVH sequences with K TTT PBS (tggtctcttcacacggac); Blue indicates HERVH sequences with F PBS; Brown indicates HERVH sequences with other PBSs (L, N, S, P); Black indicates HERVH sequences with H PBS.

4.4 HERV ORFs analyses

During evolution, the retroviral elements have been extensively damaged by simple point mutations up to large deletions or insertions, thus complicating the identification of the original genes and their original protein sequences. Some genome-wide surveys have been focused on an *in silico* reconstruction of retroviral ORFs (Martins & Villesen, 2011; Villesen et al., 2004) or, more specifically, on recovering Env genes with protein coding potential (de Parseval & Heidmann, 2005; de Parseval et al., 2003; Jern et al., 2004).

An important feature of the RetroTector program is the attempt to reconstruct the putative retroviral proteins (puteins) sequences from the different reading frames of the viral genes, thus recreating the most probable longest ORFs. The RetroTector program takes into account possible insertions and deletions, as well as the presence and frequencies of stop codons and frameshift mutations for each putein, thus estimating the ORF intactness and allowing inference of the overall protein coding potential of the identified HERV.

According to this procedure, a total number of 2272 Gag, 1743 Pro, 2843 Pol and 969 Env ORFs could be predicted within the entire set of the 3290 identified HERV sequences, that showed a range of quite varied intactness (Table 4.4). Then, these data were further filtered in order to proceed with the full characterization of the HERV sequences classified as canonical .

First, the frequency of ORF distribution, among the clearly classified HERV clades, was determined (Fig. 4.2) and compared with the whole set of predicted puteins among the 3290 HERV sequences.

Table 4.4 Predicted HERV ORFs in GRCh37/hg19

ORFs	[a]Number of reconstructed ORF	[b]Range of intactness	[c]ORF size (aa)
Gag	2272	0-73	1819-163
Pro	1743	0-11	371-25
Pol	2843	0-80	2318-73
Env	969	0-47	1506-89

[a]The number of ORFs, as reconstructed by RetroTector, was determined in all 3290 HERV sequences identified in GRCh37/hg19 (see main text for details); [b]stop codon plus frameshift (f.s.) range; [c]ORF size range in amino acid/codon.

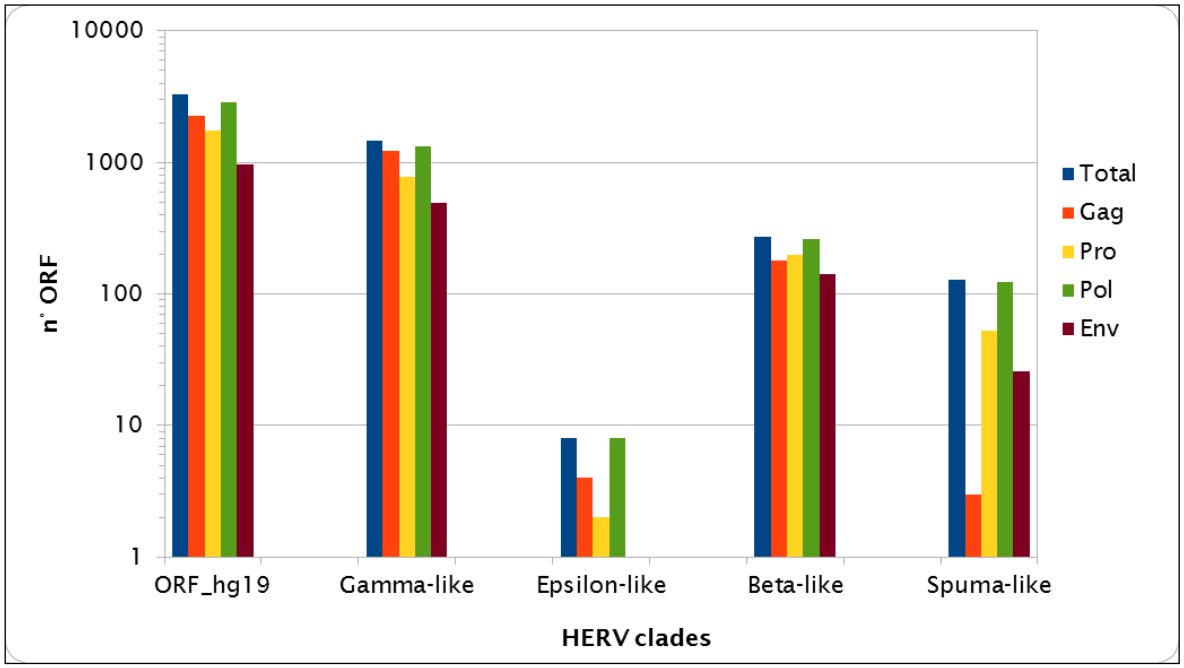

Fig. 4.2 ORFs. The distribution of the 4 viral ORF is showed for the retroviral sequences canonically classified as class I-III HERV and compared with the total number of reconstructed ORF in the entire HERV dataset (ORF_hg19).

Second, in order to determined which of the retroviral sequences harbored the most intact ORFs, the HERV clades were investigated for the lowest average of stop codons and frameshifts occurring in all ORFs. The results are reported in Table 4.5 and show that HERVFC, especially its member

HERVFc1, and HML2 are among the most intact Gamma-like and Beta-like retroviruses, respectively. The low-copy number group, HERVFC/HERVFc1, was previously described as one of the most complete HERV (Bénit et al., 2001; Bénit et al., 2003; Jern & Blomberg, 2004) and it has been implicated in the multiple sclerosis etiology in Scandinavian patients (Nexø et al., 2011; Nissen et al., 2013). A similar observation could also be done for the human-specific HML2 elements, which were proved to be among the most recent and complete endogenous retroviruses that invaded the human genome (Subramanian et al., 2011).

Table 4.5 Canonical HERV clades with the most intact ORFs

HERV clade	N° of sequences	[a]ORF intactness			
		Gag	Pro	Pol	Env
Gamma-like		n=1221	n=775	n=1336	n=497
HERVFc1	1	5±0	1±0		
[b]HERVFC	3	5.7±4	0	14±9	0
HERVFB	14		0.5±0.76		
HERV19	2		0.5±0.5		
LTR46	2		0		
HERVRB	4				2±0
ERV1_cow	1	4±0			
MER84	1	4±0			
Beta-like		n=179	n=198	n=260	n=143
[c]HML2	43	3.4±3.3		5.7±4.5	4.6±4.4
HML8	38			7±4	
Spuma-like		n=3	n=53	n=123	n=26
HERVL18	1		0.5±0.5		

[a]Values represent the lowest averages and standard deviations of stop codons plus frameshift (f.s.) per each ORF within the given HERV clade; [b]among the HERVFC sequences, the one found in chromosome X showed the most intact ORFs pattern of stop codons + f.s.: Gag=1, Pro=0, Pol=2 and Env=0; [c]some HML2 showed a similar (or better) ORF pattern intactness as HERVFC.

However, some odd data emerged from the total ORF analyses as the presence of an almost intact Gag both in ERV1_cow and MER84. ERV1_cow was phylogenetically classified as a member of the HERVRB clade that also retained some almost intact Env ORF. MER84 is a low copy number element that clustered within the HEPSI proviruses (see Chapter 3), a group of defective retroviral sequences. In this cases, the predicted Gag is short

which may have lead RetroTector to underestimate its defectiveness.

It is also worth to note that within the total number of predicted ORFs, only 30 Gag, 1024 Pro, 13 Pol and 33 Env could be really considered as open or nearly open reading frames, meaning that the sum of stop codons and frameshifts does not exceed the values of “0” and “1”, respectively. Besides the above-mentioned HERVFC and HML2, some members of the HERVH, HERVW and HERVT clades were found to have, at least, one open of the four frames. In general, these open or near open ORFs could be ascribed mainly to the canonical classified HERV clades, with the exception of 332 Pro ORF that were also found in some of the less clearly classified retroviruses. No proviral sequence out of the 3290 HERV so far investigated could be detected with all the four puteins in a completely open form. This could be seem odd since the identification of the insertional polymorphism and the full-length coding sequence HERVK113, a described member of the HERVK(HML2) group (Turner et al., 2001; Beimforde et al., 2008). However, it is noteworthy that HERVK113 is not annotated in the GRCh37/hg19 used in this project, and it could be only detected by specifically searching the human genome assembly with the HERVK113 flanking sequences (LTR) (Subramanian et al., 2011).

Finally, a particular emphasis was placed to study one of the mechanisms that underlie the retroviral gene inactivation. A statistical analysis was performed in order to define which portion of the retroviral genes was more frequently affected by stop codon mutations. In fact, even if severely damaged by an abnormal presence of stop codon mutations, HERV may still express their protein in truncated forms that could interfere with the human immune system (Nelson et al., 2003).

As shown in Fig. 4.3 the four reconstructed ORFs from the 3290 HERV sequences were statistically divided in tenths (bins), regardless of their length, and the frequency of stop codon mutations was then calculated in each bin. The results clearly demonstrated that the Gag ORF is enriched in stop codons, particularly in its second tenth (bin). Moreover, the alignment of the proviral chains from the second tenth of Gag showed that the Trp

(W), Gln (Q) and Arg (R) aminoacids are more affected than others by stop codon mutations (Fig. 4.4). The most common and ancient integrated HERV groups, such as the HERVH, HERVW and ERV9 are more prone to be affected by this mode of ORF inactivation. Thus, it could also be possible to correlate the stop codon position over the age of retroviruses integration with the aim of better understand if and how the newly integrated HERV could be inactivated by host defense mechanism.

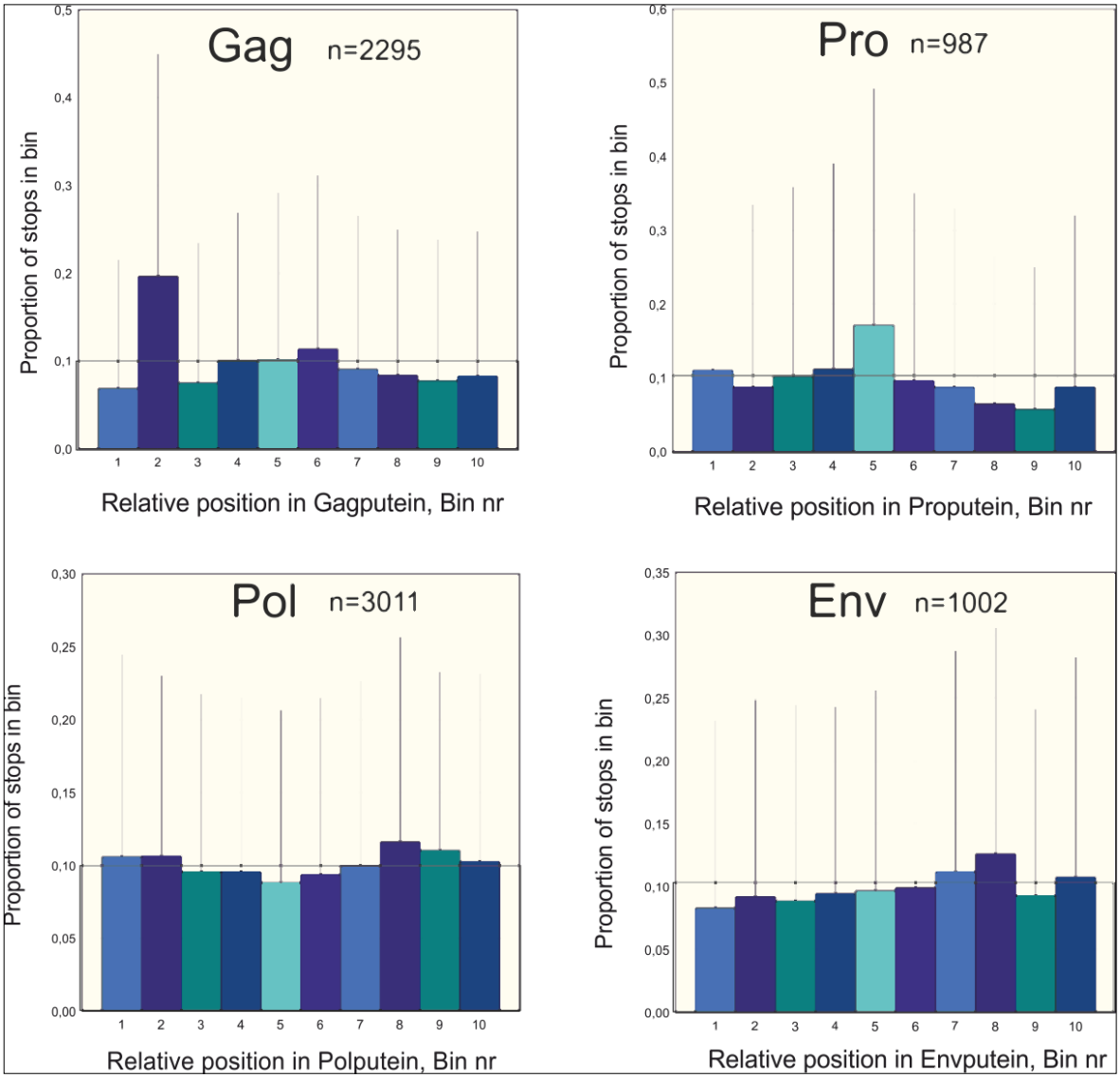

Fig. 4.3 Statistic analysis of ORFs stop codon mutations. The four reconstructed retroviral genes from the 3290 HERV were divided in bin, regardless of their length, and the frequency of stop codon mutations was calculated for each bin.

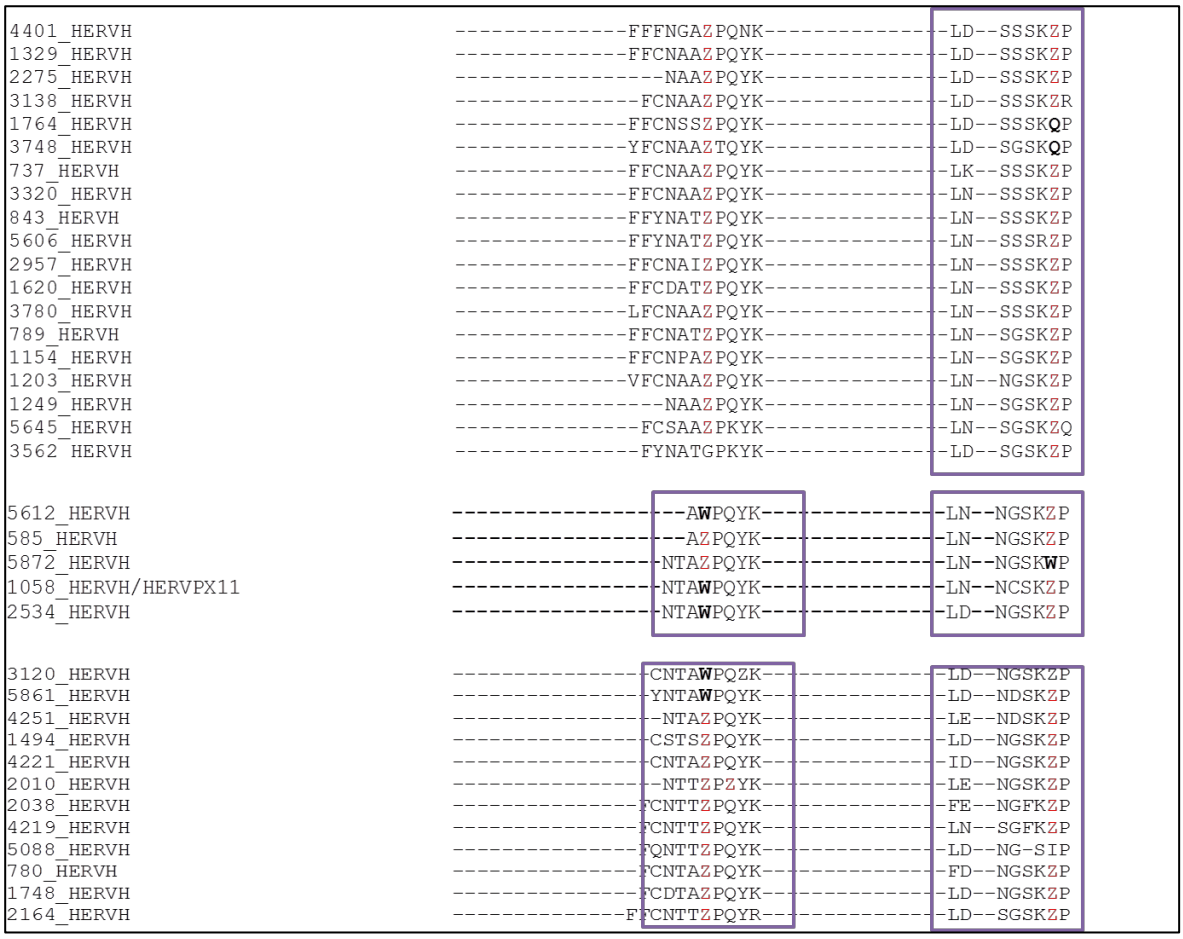

Fig. 4.4 Clustal alignment of HERV sequences from the second tenth (bin) of Gag puteins. A representative portion of the alignment showing the tryptophan (W) and glutamine (Q) amino acids affected by stop codon mutation (Z in red).

In summary, the HERV analyses, so far described, were directed mainly to define the finest genetic structure of the identified proviral sequences. As mentioned above, a further advancement in the HERV characterization is the evaluation of their background of integrations, a conditio sine qua non to study the determinants of any possible patho-physiological role of these HERV sequences in the human genome.

4.5 HERV integration clusters

As mentioned above, the second aim of this project was to define the "Where" of the identified HERV sequences. Indeed, previously reported analyses showed the occurrence of HERV integrations in chromosomes 4, 19, X and Y within the human genome assembly NCBI34/hg16 (2003 release) (Villesen, et al., 2004) or in chromosomes Y and part of chromosomes 19 within the build 31 of the human genome (Katzourakis,

et al., 2007). In order to determine if a similar trend was also true for all the 3290 identified HERV, the whole set of the 22 + Y and X chromosomes of GRCh37/hg19 were statistically analyzed for the presence, if any, of regions enriched in cluster of HERV integrations.

First, each chromosome was sliced in 10 million-bases bins (fragments) and the Start/End positions of the 3290 HERV sequences were mapped back onto each of them generating the graphical distribution reported in Fig.4.5.

The observed HERV positions were scored against a random distribution of the total number of detected HERV on the 3.2 Gigabases of the whole genome GRCh37/hg19. As shown in Fig. 4.5, a total number of 9 clusters distributed in chromosomes 3, 4, 8, 19, X and Y were found, thus partially confirming the results reported by Villesen et al. The overall frequency of these clusters represented 13% of the whole genome HERV integrations, the largest one being present in the third bin of chromosome Y with 85 HERV integrations.

Next, supported by the previously performed classification, the clusters of HERV integrations were scrutinized for any genus and/or retroviral clade correlation. As reported in Fig. 4.6, most of the HERV integrations found in all clusters belong to the Gamma-like retroviruses with a particular over-representation, with respect to the whole genome frequency, into specific bins of chromosomes 3 and X. It is also worth to note that few (13) sequences were identified as ALU or LINE, whereas most of the retroviral clusters of integrations could be referred to the canonically classified HERV clades. Moreover, the integration patterns analysis showed a tendency for proviruses from the same clade, such as HERVH, HERVE and ERV9 (ERV9 sequences were most represented in the second and third bins of chromosome Y), to occur together, within 100000 bases. This could be possibly explained with local retroviral sequence duplications events.

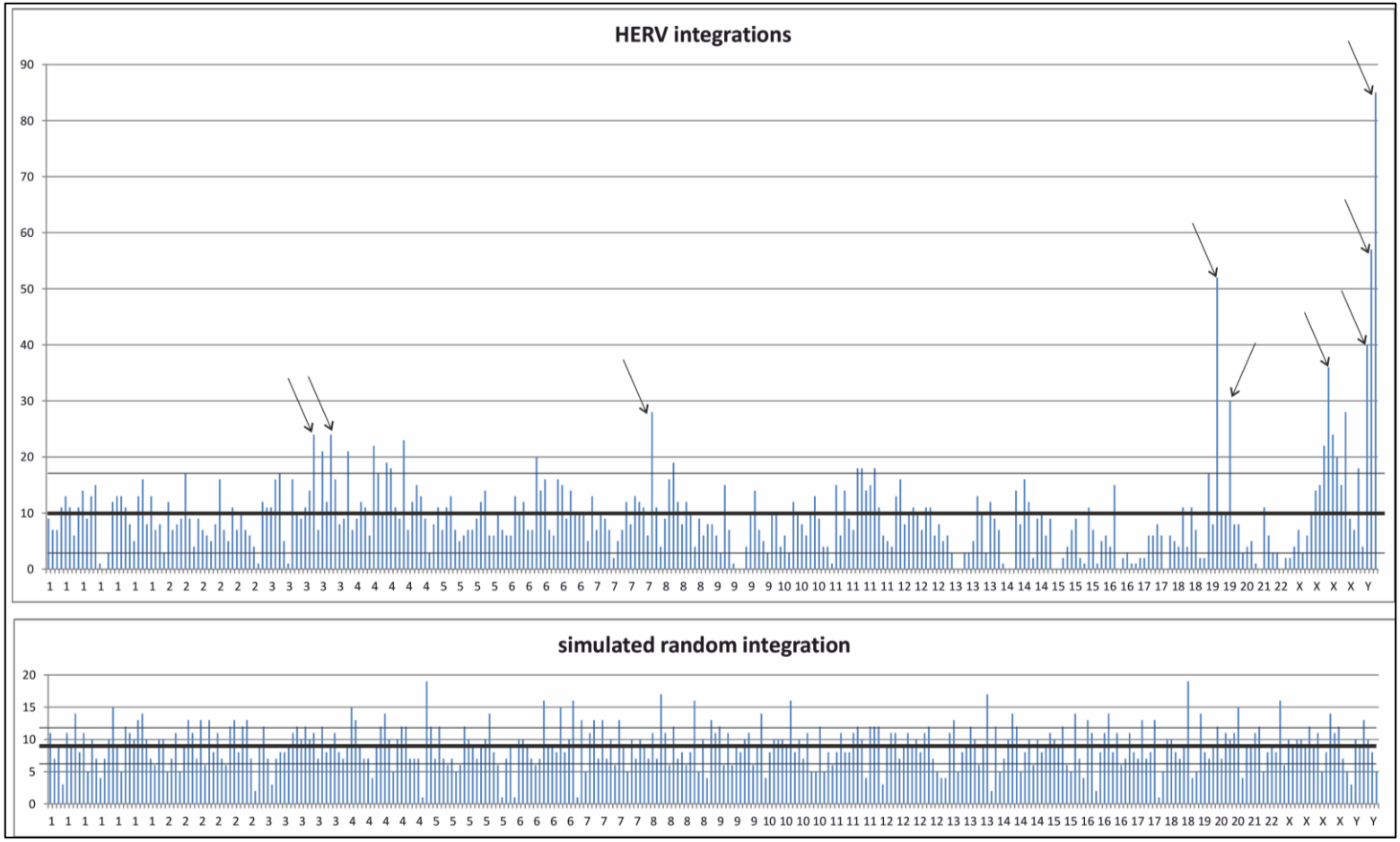

Fig.4.5 Statistic analysis of HERV clustering in GRCh37/hg19. Each chromosome was divided in bin (10 million-bases) and the average of real (10±8) vs the of random (9±3) HERV integrations was calculated. The arrows indicated the HERV clusters in chr 3, 4, 8, 19, X and Y.

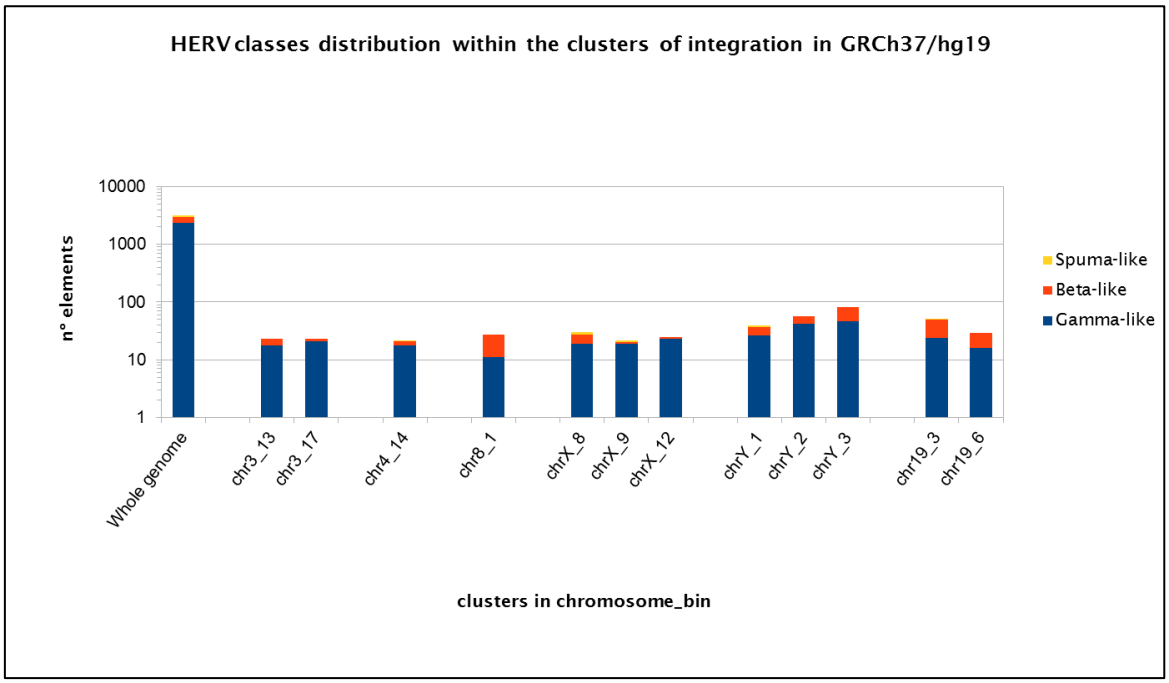

Fig. 4.6 Most frequent HERV classes found in clusters of integration. A general overrepresentation of Gamma-like sequences with respect to the whole genome frequencies was found among all clusters. Some HERV clades (i.e. HERVH) were more represented than others (see details in Chapter 4.5).

Finally, the general picture of HERV background of integrations could be completed with the localization of the retroviral sequences with respect to their proximity to the transcriptional units (coding and non-coding genes) annotated in the human genome.

4.6 The "human genes" context of HERV integrations

Previous broad characterizations of the context of genomic integrations of transposable elements and retroelements, also encompassing HERV, were both performed in relation to coding (Medstrand et al., 2005; Medstrand et al., 2002; Villesen et al., 2004; Pérot et al., 2012) and non-coding human genes (Kapusta et al., 2013; Kelley & Rinn, 2012). A similar evaluation of the 3290 classified HERV sequences was then performed with respect to selected datasets of transcriptional units (TUx) with the aim to catch, if any, an association between HERV genus/clade and their relative distances from the TUx.

The chromosomal coordinates of the 3290 HERV sequences were matched with the transcript start and end positions of a set of annotated genes in GRCh37/hg19, the Ensembl dataset (release 71). Using an algorithm (kindly performed by J. Blomberg), each chromosome was divided into bins (fragments) and the minimum distances from each HERV sequence within the start and end positions of the TUx and in the range of 0-1, 1-10, 10-20 up to >90 kb (kilobases) both upstream and downstream the TUx were calculated. The results of this analysis are plotted in Fig. 4.7 and clearly highlighted the tendency for HERV sequences to integrate preferentially, both upstream and downstream, within 10 kb from TUx. Moreover, members of the HERVH clade were more prone than others to integrate near the TUx 5'-end and in antisense orientation. These TUx were further investigated for their gene annotation in Gene Ontology. A few of them turned out to be associated to protein products, whereas most were identified as long non-coding RNA (lnc-RNA). These data are of particular interest if it is considered the predominance of non-coding RNA within the human transcriptome and, even more, the functional role in the human genome regulation (Mattick & Makunin, 2006; Mercer et al., 2009).

Then, a similar procedure was also applied to compare TUx datasets from different sources, that is Vega 51, a finest collection of annotated and manually curated genes available at Ensembl genome browser, and the GENCODE collection (Derrien et al., 2012; Harrow et al., 2012) of manually curated long non-coding RNA (lnc). These preliminary data are illustrated in Fig. 4.8 where it is noteworthy that the proportion of long distances of integrations to TUx is greater for long non-coding RNA than for coding RNA (Ensembl 71 and Vega 51).

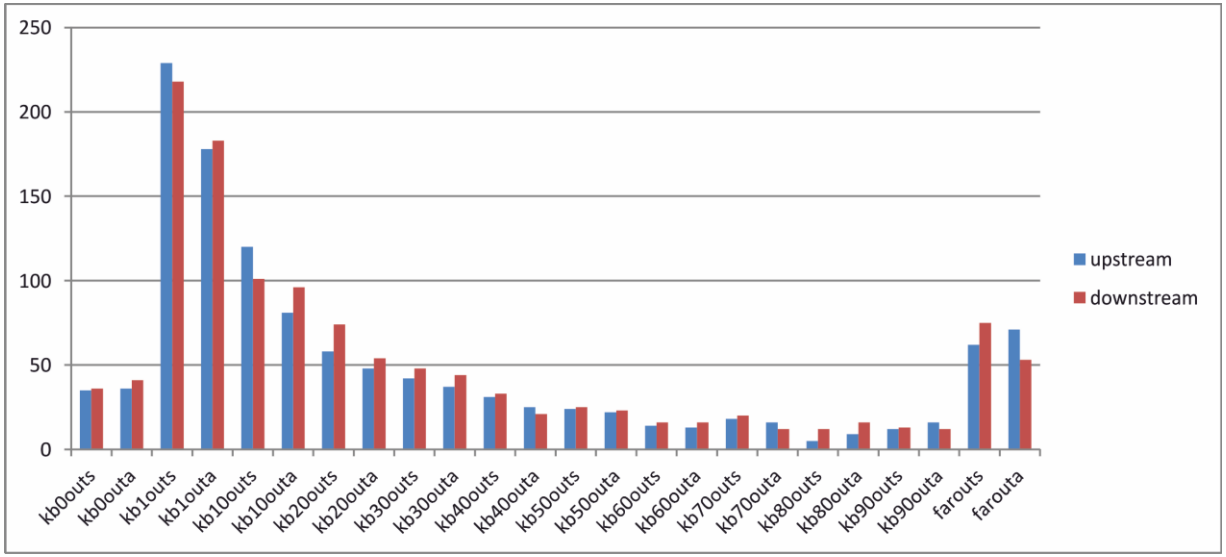

Fig. 4.7 Distribution of HERV sequences *vs* Ensembl 71 Tux dataset. HERV sequences were investigated both for their distance (x axis: in kilobases), outside (out) upstream and downstream position and relative orientation (s: sense, a: antisense) with respect to TUx.

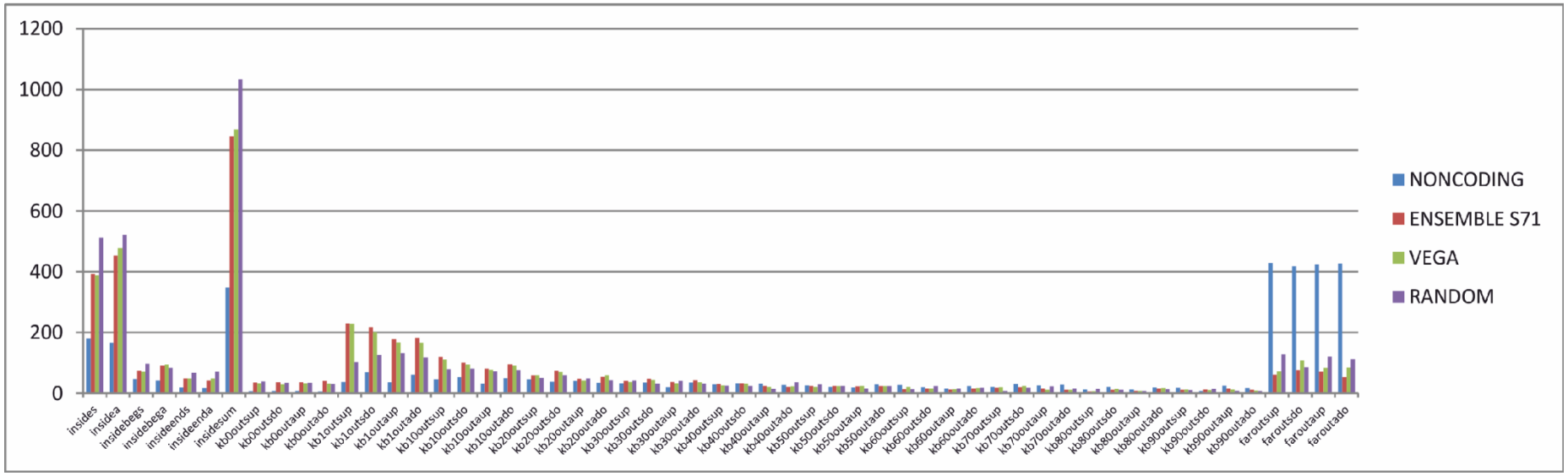

Fig. 4.8 Distribution of HERV sequences *vs* different Tux dataset. HERV sequences were investigated for: i) distance (x axis: in kilobases), ii) inside position: near 5'-end (beg) and 3'-end (end), iii) outside position: upstream (up) and downstream (do) and iv) relative orientation (s: sense, a: antisense) with respect to TUx.

4.7 Discussion

"Know your enemy, know yourself", this ancestral sentence could be easily moved to the HERV world. Any progress in understanding the complex role that these inherited sequences could have in "helping" or "fighting" the human genome could not be addressed if, first of all, HERV remained not clearly defined.

Starting from a model-based search algorithm, RetroTector, 3290 HERV proviral sequences were identified and classified in the human genome reference GRCh37/hg19.

The finest definition of the HERV genetic structure was here extended on the investigation of a series of structural markers that overall contributed both to confirm or improve, such as in the case of new dUTPase^{Pol} and PBS signatures in the HERVL clades, the previous classification.

The analysis of retroviral ORF stop codon mutations could also contribute, on the one hand, to identify the full - or partial - protein coding potential of the HERV sequences supporting the *in vitro* design of tests to assay the retroviral protein expression, and, on the other hand, to better define the mode of inactivation of newly integrated retroviruses.

Finally, the overall, even if not exhaustive, survey of the genomic neighborhood of HERV integrations revealed that some gammaretroviral clades preferentially integrate within certain chromosomes and next to promoters or inside long non coding-RNA. These results show that HERV sequences *cis*-effects are to be expected.

In conclusion, the exhaustive classification and the wide characterization of the whole 3290 HERV sequences found in the GRCh37/hg19 led to the development of a "collection/database" that could be a useful starting point for upcoming studies of HERV.

5. Conclusions

The study of HERVs represents an intriguing challenge. After 30 years of extensive research in this field, some basic questions regarding the HERV classification, structure and role in modulating the human pathophysiology are still unsolved. An advance in the HERV knowledge must include a clear definition of the exact genetic structure of these retroviral sequences.

The main objective of this project was the identification and characterization of the most intact HERV sequences present in the human genome assembly GRCh37/hg19. HERV identification was performed through a bioinformatics approach using the RetroTector software and a total number of 3290 proviral sequences actually integrated within the human genome were found and further characterized.

The main task was to achieve a complete classification of the HERV sequences; therefore a complex procedure involving the Simage (similarity image) analysis was developed. Simage analysis led to the taxonomic identification of about 95% of the retroviral sequences and to the description of a final number of 40 HERV clades that, partially, overlapped previously reported HERV groups (also called "families") (Mager & Medstrand, 2005; Katzourakis & Tristem, 2005a). Possibly, some observed differences could be explained with the methodologies applied for both the identification and the classification of HERV sequences. Indeed, the focus of this project was to define, as far as possible, a precise number of almost integer HERV proviruses, thus discarding the enumeration of solo-LTRs not clearly identified by RetroTector. Moreover, the complex phylogenetic analysis, mainly based on Simage, allowed to better define the "borderline" between some groups that have been previously listed as separate clades (i.e. HERV9 and HERV30), and also to introduce a new HEPSI clade within the Class I HERV or to identify short stretches of Errantivirus-like similarity within the Pol regions of some HERV proviruses. Strangely, no Gag-like sequences of Errantivirus origin were identified (Campillos et al., 2007) possibly because of some RetroTector detection limits.

Simage analysis also contributed to define the complex genetic structure of most HERV sequences and determined the presence of a high number of mosaic structures, possible recombinant forms, with a level of detail not previously appreciated. In fact, the most extensive descriptions of HERV recombination events referred to the homologous recombination that is responsible for the solo-LTR formation (Belshaw et al., 2007; Katzourakis et al, 2007) or for the documented intra-chromosomal recombination between two homologous HERV15 (Repbase identifier for the RRHERVI group) sequences located in chromosome Y that seemed to be responsible for male infertility because of the Azoospermia factor a (AZFa) microdeletion (Kamp et al., 2000). Nonetheless, an overall description and even enumeration of "mosaicisms" occurring within HERV internal structures was not listed. The results emerged from Simage analysis of the 3290 HERV proviruses could contribute to improve the previous knowledge about the chimeric nature of endogenous retroviral sequences.

The finest characterization of the HERV sequences was then extended to the investigation of a series of structural markers. Among them, the PBS sequence analysis highlighted unexpected PBS usage patterns within some HERV clades such as HERVH and ERV3.

The protein coding potential of the entire set of HERV sequences was also determined by the analysis of the main reconstructed ORFs. A similar genome-wide survey of retroviral ORFs in the human genome (Villesen et al., 2004) has highlighted the presence of 59 longest ORFs (Gag, Pol and Env) with a length size (from stop codon to stop codon) of > 500 amino acids (for Gag and Env) and >700 (for Pol) with a Betaretroviral classification for most of them. These data could be partially compared to the 76 Gag, Pol, and Env ORFs, here identified, defined as open or near open (a 0-1 range of stop codon plus frame shifts), with 52 out of 76 from Beta-like sequences. However, some differences emerged when ORFs size length is compared between the two sets. In fact, the length range for the 76 ORFs is comprised within 450-720, 145-850 and 600-615 amino acids, for Gag, Pol and Env, respectively, while the range size described by Villesen et al. were 500 - >1000 for all their longest ORFs.

The different methods and cutoffs applied for the definition of ORFs intactness could be also responsible for the identification of different numbers of Gag and Env (30 and 33, respectively) versus the 17 and 29 identified by Villesen et al.

Finally, the ORFs potential of the classified 3290 HERV sequences was reinforced with the statistical analysis of stop codon mutation. Evidence showed that the most common and ancient integrated HERV sequences were more prone to be affected by a high frequency of nonsense mutations in their Gag ORF.

To complete the HERV characterization, a preliminary survey of the genomic neighborhood of HERV integrations was performed. According to previously reported data (Villesen et al., 2004; Kazourakis et al, 2007), a strong preference for HERV proviruses to integrate in chromosomes 4, 19, X and Y was confirmed and further extended to chromosomes 3 and 8 that also showed peaks of retroviral sequences integrations mainly belonging to Gamma-like clades.

The pattern of integration of the 3290 proviruses towards transcription units (TUx or genes) was also investigated. Indeed, a previously observed profile has showed the tendency for Class I-III HERV to be underrepresented inside genes especially if HERV sequences were in the same gene orientation and a trend of underrepresentation both for Class I and III has been showed within a 5 kilobases window from genes (Medstrand et a al., 2002) both upstream and downstream (for Class III) or only downstream (for Class I). The integration pattern resulting from the HERV sequences here studied, revealed a similar underrepresentation of elements within the beginning of genes and a marked antisense orientation, thus confirming the data reported by Medstrand et al. and a tendency for most HERV sequences (regardless of the HERV classifications) to integrate within 10 kb from TUx (both upstream and downstream). It is noteworthy that this window range is slightly different from the more recently proposed 8 kb zone showing a low gene sense orientation density upstream of a dedicated set of LTRs with specific promoter functions

(Pérot et al., 2012). However, the 10 kb zone here calculated is associated to the ends positions of the entire proviral structures with respect to the TUx, whereas the Pérot et al. 8 kb interval is derived from a dataset of 1232 LTRs (mainly solo-LTRs or provirus-associated) for which a specific function (promoter, poly-A signal or silent LTR) has been newly identified. It is hence reasonable to assume that the different result could be derived from experimental differences.

In conclusion, a robust and innovative HERV classification procedure was performed with the aim to contribute to a clear definition of the finest genetic structure of most retroviral sequences.

The entire set of HERVs and the main features emerged from their overall analysis were used to develop the most updated database of HERV proviral sequences found in the human genome.

6. Bibliography

Andersson, M. L., Lindeskog, M., Medstrand, P., Westley, B., May, F., & Blomberg, J. (1999). Diversity of human endogenous retrovirus class II-like sequences. *The Journal of General Virology*, *80 (Pt 1)*, 255–60

Andersson, A.-C., Yun, Z., Sperber, G. O., Larsson, E., & Blomberg, J. (2005). ERV3 and Related Sequences in Humans : Structure and RNA Expression. *Journal of Virology*, *79*(14), 9270.

Antony, J. M., Deslauriers, A. M., Bhat, R. K., Ellestad, K. K., & Power, C. (2011). Human endogenous retroviruses and multiple sclerosis: innocent bystanders or disease determinants? *Biochimica et Biophysica Acta*, *1812*(2), 162–76.

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

Baillie, G. J., Lagemaat, L. N. Van De, Baust, C., & Mager, D. L. (2004). Multiple Groups of Endogenous Betaretroviruses in Mice , Rats , and Other Mammals. *Journal of Virology*, *78*(11), 5784–5798.

Balada, E., Ordi-Ros, J., & Vilardell-Tarrés, M. (2009). Molecular mechanisms mediated by human endogenous retroviruses (HERVs) in autoimmunity. In *Reviews in medical virology* (Vol. 19, pp. 273–286). Wiley Online Library.

Balada, E., Vilardell-Tarrés, M., & Ordi-Ros, J. (2010). Implication of human endogenous retroviruses in the development of autoimmune diseases. *International Reviews of Immunology*, *29*(4), 351–70.

Bannert, N., & Kurth, R. (2006). The evolutionary dynamics of human endogenous retroviral families. *Annual Review of Genomics and Human Genetics*, *7*, 149–73.

Beimforde, N., Hanke, K., Ammar, I., Kurth, R., & Bannert, N. (2008).

Molecular cloning and functional characterization of the human endogenous retrovirus K113. *Virology*, 371(1), 216–25.

Belshaw, R., Watson, J., Katzourakis, A., Howe, A., Woolven-Allen, J., Burt, A., & Tristem, M. (2007). Rate of recombinational deletion among human endogenous retroviruses. *Journal of virology*, *81*(17), 9437–42.

Benachenhou, F., Jern, P., Oja, M., Sperber, G. O., Blikstad, V., Somervuo, P., … & Blomberg, J. (2009). Evolutionary conservation of orthoretroviral long terminal repeats (LTRs) and ab initio detection of single LTRs in genomic data. *PloS One*, *4*(4), 5179.

Bénit, L., De Parseval, N., Casella, J. F., Callebaut, I., Cordonnier, A., & Heidmann, T. (1997). Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. *Journal of Virology*, 71(7), 5652–7.

Bjerregaard, B., Holck, S., Christensen, I. J., & Larsson, L.-I. (2006). Syncytin is involved in breast cancer-endothelial cell fusions. Cellular and Molecular Life Sciences : CMLS, 63(16), 1906–11.

Blaise, S., de Parseval, N., Bénit, L., & Heidmann, T. (2003). Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proceedings of the National Academy of Sciences of the United States of America, 100(22), 13013–8.

Blikstad, V., Benachenhou, F., Sperber, G. O., & Blomberg, J. (2008). Evolution of human endogenous retroviral sequences: a conceptual account. *Cellular and Molecular Life Sciences : CMLS*, *65*(21), 3348–65.

Blomberg, J., Benachenhou, F., Blikstad, V., Sperber, G. O., & Mayer, J. (2009). Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. *Gene*, *448*(2), 115–23.

Blomberg, J., Sperber, G. O., Jern, P., & Benachenhou, F. (2010). Towards a

retrovirus database , RetroBank. *Medical Biochemistry*, 19–22.

Blond, J. L., Beseme, F., Duret, L., Bouton, O., Bedin, F., Perron, H., …& Mallet, F. (1999). Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family. *Journal of Virology*, 73(2), 1175–85.

Boeke JD & Stoye JP (1997) Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements. In Coffin JM, Hughes SH, Varmus HE (ed.), *Retroviruses*, Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.

Boese, a, Sauter, M., Galli, U., Best, B., Herbst, H., Mayer, J., … Mueller-Lantzsch, N. (2000). Human endogenous retrovirus protein cORF supports cell transformation and associates with the promyelocytic leukemia zinc finger protein. Oncogene, 19(38), 4328–36.

Bowzard, J. B., Bennett, R. P., Krishna, N. K., Ernst, S. M., Rein, a, & Wills, J. W. (1998). Importance of basic residues in the nucleocapsid sequence for retrovirus Gag assembly and complementation rescue. *Journal of Virology*, 72(11), 9034–44.

Brodziak, A., Ziolko, E., Muc-Wierzgon, E., Nowakowska-Zajdel, E., Kokot, T., & Klakla, K. (2012). The role of human endogenous retroviruses in the pathogenesis of autoimmune diseases. *Medical Science Monitor*, *18*(6).

Buzdin, A. (2007). Human-specific endogenous retroviruses. *TheScientificWorldJournal*, *7*, 1848–68.

Campillos, M., Doerks, T., Shah, P. K., & Bork, P. (2006). Computational characterization of multiple Gag-like human proteins. *Trends in Genetics : TIG*, *22*(11), 585–589.

Cegolon, L., Salata, C., Weiderpass, E., Vineis, P., Palù, G., & Mastrangelo, G. (2013). Human endogenous retroviruses and cancer prevention: evidence and prospects. *BMC Cancer*, *13*, 4.

Coffin JM. (1997). Retroviruses. Cold Spring Harbor (NY): Cold Spring

Harbor Laboratory Press; Hughes SH, Varmus HE, editors.

Cohen, C. J., Lock, W. M., & Mager, D. L. (2009). Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene*, *448*(2), 105–14.

Copeland, N. G., Hutchison, K. W., & Jenkins, N. a. (1983). Excision of the DBA ecotropic provirus in dilute coat-color revertants of mice occurs by homologous recombination involving the viral LTRs. *Cell*, *33*(2), 379–87.

Cordonnier, A., Casella, J. F., & Heidmann, T. (1995). Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *Journal of virology*, *69*(9), 5890–97.

Deininger, P. L., & Batzer, M. a. (2002). Mammalian retroelements. *Genome Research*, *12*(10), 1455–65.

Denne, M., Sauter, M., Armbruester, V., Licht, J. D., Roemer, K., & Mueller-Lantzsch, N. (2007). Physical and functional interactions of human endogenous retrovirus proteins Np9 and rec with the promyelocytic leukemia zinc finger protein. *Journal of Virology*, *81*(11), 5607–16.

de Parseval, N., Lazar, V., Casella, J. F., Bénit, L., & Heidmann, T. (2003). Survey of Human Genes of Retroviral Origin : Identification and Transcriptome of the Genes with Coding Capacity for Complete Envelope Proteins. *Journal of Virology*, *77*(19), 10414–10422.

de Parseval, N., & Heidmann, T. (2005). Human endogenous retroviruses: from infectious elements to human genes. *Cytogenetic and Genome Research*, 110(1-4), 318–32.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., … & Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research*, *22*(9), 1775–89.

Eickbush, T. H., & Jamburuthugoda, V. K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus*

research, *134*(585), 221–234.

Enssle, J., Jordan, I., Mauer, B., & Rethwilm A. (1996). Foamy virus reverse transcriptase is expressed independently from the Gag protein. *Proceedings of the National Academy of Sciences of the United States of America*, 93(9), 4137–41.

Flockerzi, A., Ruggieri, A., Frank, O., Sauter, M., Maldener, E., Kopper, B., … & Mayer, J. (2008). Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project. *BMC Genomics*, 9, 354.

Flugel, R. M. (1991). Spumaviruses: a group of complex retroviruses. *Journal of Acquired Immune Deficiency Syndromes*, 4, 739–750.

Freed, E. O. (2002). Viral Late Domains. *Journal of Virology*, 76(10), 4679–4687.

Galli, U. M., M. Sauter, B. Lecher, S. Maurer, H. Herbst, K. Roemer, and N. Mueller-Lantzsch. (2005.) Human endogenous retrovirus rec interferes with germ cell development in mice and may cause carcinoma in situ, the predecessor lesion of germ cell tumors. *Oncogene* 24:3223–3228.

Gifford, R. J., & Tristem, M. (2003). The Evolution , Distribution and Diversity of Endogenous Retroviruses. *Virus Genes*.

Goff SP (2007). Retroviridae: the retroviruses and their replication. In Knipe DM, Howley PM (ed.), *Fields virology,* 5th ed., Lippincott Williams and Wilkins Philadelphia, PA.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … & Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–74.

Hu, W. S., & Temin, H. M. (1990). Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of

genetic recombination. Proceedings of the National Academy of Sciences of the United States of America, 87(4), 1556–60.

Jern, P., Sperber, G. O., & Blomberg, J. (2004). Definition and variation of human endogenous retrovirus H. *Virology*, 327(1), 93–110

Jern, P., & Sperber, G. O. (2005a). Sequence Variability , Gene Structure , and Expression of Full-Length Human Endogenous Retrovirus H. *Journal of Virology*, *79*(10), 6325–6337.

Jern, P., Sperber, G. O., & Blomberg, J. (2005b). Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. *Retrovirology*, *2*, 50.

Jern, P., & Coffin, J. M. (2008). Effects of retroviruses on host genome function. *Annual Review of Genetics*, *42*, 709–32.

Jühling, F., Mörl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F., & Pütz, J. (2009). tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Research*, *37*(Database issue), D159–62.

Jurka, J., Kapitonov, V. V, Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*(1-4), 462–7.

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., … & Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genetics*, 9(4), e1003470

Katzourakis, A. & Tristem, M. (2005a). Phylogeny of Human Endogenous and Exogenous Retroviruses In Sverdlov E. (ed) Retroviruses and Primate Retroviruses and Primate Genome Evolution. *Eurekah/Landes Bioscience*.

Katzourakis, A., Rambaut, A., & Pybus, O. G. (2005b). The evolutionary dynamics of endogenous retroviruses. *Trends in Microbiology*, *13*(10), 463–8.

Katzourakis, A., Pereira, V., & Tristem, M. (2007). Effects of recombination rate on human endogenous retrovirus fixation and persistence. *Journal of Virology*, *81*(19), 10712–7.

Kelley, D., & Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology*, 13(11), R107

Kurth, R., & Bannert, N. (2010). Beneficial and detrimental effects of human endogenous retroviruses. International Journal of Cancer. *International Journal of Cancer*, 126(2), 306–14

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., …& Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Lee, E.-G., Stenbak, C. R., & Linial, M. L. (2013). Foamy virus assembly with emphasis on pol encapsidation. *Viruses*, 5(3), 886–900

Leib-Mösch, C., Seifarth, W., & Schön, U. (2005). Influence of Human Endogenous Retroviruses on Cellular Gene Expression. In Retroviruses and Primate Genome Evolution (pp. 123–143).

Lerat, E. (2010). Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity, 104(6), 520–33.

Lole, K., & Ray, SC (1999) Full-Length Human Immunodeficiency Virus Type 1 Genomes from Subtype C-Infected Seroconverters in India, with Evidence of Intersubtype Recombination. Journal of virology, 73, 152-160.

Lower, R., Lower, J., & Kurth, R. (1996). The viruses in all of us : Characteristics and biological significance of human endogenous retrovirus sequences. Proceedings of the National Academy of Sciences, 93(May), 5177–5184.

Lynch, C., & Tristem, M. (2003). A Co-opted gypsy -type LTR-Retrotransposon Is Conserved in the Genomes of Humans , Sheep , Mice , and Rats. *Current biology : CB*, *13*, 1518–1523.

Mager, D. L., & Medstrand, P. (2003). Retroviral Repeat Sequences. In *Encyclopedia of human genome* (pp. 1–7).

Magiorkinis, G., Belshaw, R., & Katzourakis, A. (2013). " There and back again ": revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era. *Philosophical Transactions of the Royal Society*, B 2013 368

Malik, H. S. (2000). Poised for Contagion: Evolutionary Origins of the Infectious Abilities of Invertebrate Retroviruses. Genome Research, 10(9), 1307–1318. doi:10.1101/gr.145000

Màllet, F., Bouton, O., Prudhomme, S., Cheynet, V., Oriol, G., Bonnaud, B., …& Mandrand, B. (2004). The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proceedings of the National Academy of Sciences of the United States of America, 101(6), 1731–6.

Martins, H., & Villesen, P. (2011). Conservation of ancient full-length open reading frames in vertebrate endogenous retroviruses. *Retrovirology*, 8(Suppl 2), P47

Mattick, J. S., & Makunin, I. V. (2006). Non-coding RNA. *Human Molecular Genetics*, 15 Spec No(1), R17–29

Mayer, J., & Meese, E. (2003). Presence of dUTPase in the various human endogenous retrovirus K (HERV-K) families. *Journal of Molecular Evolution*, 57, 642–649

Mayer, J., Blomberg, J., & Seal, R. L. (2011). A revised nomenclature for transcribed human endogenous retroviral loci. *Mobile DNA*, *2*(1), 7.

Medstrand, P., Landry, J. R., & Mager, D. L. (2001). Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. *The Journal of Biological Chemistry*, *276*(3), 1896–903.

Medstrand, P., van de Lagemaat, L. N., & Mager, D. L. (2002). Retroelement

distributions in the human genome: variations associated with age and proximity to genes. *Genome Research*, 12(10), 1483–95

Medstrand, P., van de Lagemaat, L., Dunn, C., Landry, J.-R., Svenback, D., & Mager, D. (2005). Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenetic and Genome Research*, 110(1-4), 342–52

Mercer, T. R., Dinger, M. E., & Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nature Reviews. Genetics*, 10, 155

Mirambeau, G., Lyonnais, S., & Gorelick, R. J. (2010). Features, processing states, and heterologous protein interactions in the modulation of the retroviral nucleocapsid protein function. *RNA Biology*, 7(6), 724–734

Nelson, P. N., Carnegie, P. R., Martin, J., Davari Ejtehadi, H., Hooley, P., Roden, D., …& Murray, P. G. (2003). Demystified. Human endogenous retroviruses. *Molecular Pathology* : MP, 56(1), 11–18

Nevins JR (2007). Cell transformation by viruses. In Knipe DM, Howley PM (ed.), *Fields virology,* 5th ed., Lippincott Williams and Wilkins Philadelphia, PA.

Nexø, B. A., Christensen, T., Frederiksen, J., Møller-Larsen, A., Oturai, A. B., Villesen, P., …& Pedersen, F. S. (2011). The etiology of multiple sclerosis: genetic evidence for the involvement of the human endogenous retrovirus HERV-Fc1. *PloS One*, 6(2), e16652

Nissen, K. K., Laska, M. J., Hansen, B., Terkelsen, T., Villesen, P., Bahrami, S., … Nexø, B. a. (2013). Endogenous retroviruses and multiple sclerosis--new pieces to the puzzle. *BMC Neurology*, 13(1), 111

Oja, M., Sperber, G. O., Blomberg, J., & Kaski, S. (2005). Self-organizing map-based discovery and visualization of human endogenous retroviral sequence groups. *International journal of neural systems*, *15*(3), 163–79.

Paces, J., Pavlícek, A., Zika, R., Kapitonov, V. V, Jurka, J., Paces, V., & Pavlicek, A. (2004). HERVd: the Human Endogenous RetroViruses Database:

update. *Nucleic acids research*, *32* (Database issue), D50.

Pérot, P., Mugnier, N., Montgiraud, C., Gimenez, J., Jaillard, M., Bonnaud, B., & Màllet, F. (2012). Microarray-based sketches of the HERV transcriptome landscape. *PloS One*, *7*(6), e40194.

Ribet, D., Louvet-Vallée, S., Harper, F., de Parseval, N., Dewannieux, M., Heidmann, O., …& Heidmann, T. (2008). Murine endogenous retrovirus MuERV-L is the progenitor of the "orphan" epsilon viruslike particles of the early mouse embryo. *Journal of virology*, *82*(3), 1622–5.

Smit, A. F. *(1996).* The origin of interspersed repeats in the human genome*. Current opinion in genetics & development 6 (6),* 743-749.

Sperber, G. O., Airola, T., Jern, P., & Blomberg, J. (2007). Automated recognition of retroviral sequences in genomic data--RetroTector. *Nucleic Acids Research*, *35*(15), 4964–76.

Stefanov, Y., Salenko, V., & Glukhov, I. (2012). Drosophila errantiviruses. *Mobile Genetic Elements*, *2*(1), 36–45.

Stoye, J. P. (2012). Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nature Reviews. Microbiology*, *10*(6), 395–406.

Stuhlmannt, H., & Berg, P. (1992). Homologous Recombination of Copackaged Retrovirus RNAs during Reverse Transcription. Journal of Virology, 66(4), 2378–2388.

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

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, *28*(10), 2731–9.

Tarlinton, R., Meers, J., & Young, P. (2008). Biology and evolution of the endogenous koala retrovirus. *Cellular and Molecular Life Sciences : CMLS*, *65*(21), 3413–21.

Tristem, M. (2000). Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *Journal of Virology*, *74*(8), 3715–30.

Turner, G., Barbulescu, M., Su, M., Jensen-Seaman, M. I., Kidd, K. K., & Lenz, J. (2001). Insertional polymorphisms of full-length endogenous retroviruses in humans. *Current Biology* : CB, 11(19), 1531–5

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

Vogt, PK (1997). Historical introduction to the general properties of Retroviruses. In Coffin JM, Hughes SH, Varmus HE (ed) *Retroviruses*. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.

Voisset, C., Weiss, R. A., & Griffiths, D. J. (2008). Human RNA "rumor" viruses: the search for novel human retroviruses in chronic disease. *Microbiology and Molecular Biology Reviews : MMBR*, *72*(1), 157–96.

Volff, J. N. (2009). Cellular genes derived from Gypsy/Ty3 Retrotransposons in Mammalian Genomes. *Natural Genetic Engineering and Natural Genome Editing*, *1178*, 133–243.

Young, G. R., Stoye, J. P., & Kassiotis, G. (2013). Are human endogenous retroviruses pathogenic? An approach to testing the hypothesis. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, 1–10