Università degli Studi di Cagliari

DOTTORATO DI RICERCA

Scienze della Vita, dell'Ambiente e del Farmaco

Ciclo XXIX

TITOLO TESI

Identification and characterization of Type W Human Endogenous Retroviruses (HERV-W) in humans and comparative analysis of the group in non-human primates: new insights on HERV-W diffusion in Simiiformes and analysis of evolutionarily highly related sequences in New World Monkeys

Settore scientifico disciplinare di afferenza

Microbiologia BIO/19

Presentata da:	Nicole Grandi
Coordinatore Dottorato	Prof. Enzo Tramontano
Tutor	Prof. Enzo Tramontano

Esame finale Anno Accademico 2015 – 2016

Tesi discussa nella sessione d'esame marzo – aprile 2017

Università degli Studi di Cagliari

DOTTORATO DI RICERCA

Scienze della Vita, dell'Ambiente e del Farmaco

Ciclo XXIX

TITOLO TESI

Identification and characterization of Type W Human Endogenous Retroviruses (HERV-W) in humans and comparative analysis of the group in non-human primates: new insights on HERV-W diffusion in Simiiformes and analysis of evolutionarily highly related sequences in New World Monkeys

Settore scientifico disciplinare di afferenza

Microbiologia BIO/19

Presentata da:	Nicole Grandi
Coordinatore Dottorato	Prof. Enzo Tramontano
Tutor	Prof. Enzo Tramontano

Esame finale Anno Accademico 2015 – 2016
Tesi discussa nella sessione d'esame marzo – aprile 2017

Summary

ABSTRACT

Human Endogenous Retroviruses (HERVs) are ancient retroviral infections relics that became stable components of the human genome, constituting ~8% of our DNA. While HERVs genomic characterization is still ongoing, impressive amounts of data have been obtained regarding some HERV groups general expression across human tissues, primarily to find a role in human pathogenesis. Among HERV groups, one of the most explored is the HERV-W one, initially studied for the presence of an element integrated in locus 7q21.2 and encoding a functional envelope protein, coopted for human placentation. The HERV-W group has been also intensively investigated for its putative role in several diseases, such as cancer, inflammation, autoimmunity and infectious disorders. However, despite the large amount of studies tentatively linking the group expression to human pathogenesis, no conclusive correlations have been demonstrated so far. In general, both i) the absence of an exhaustive characterization of the HERV-W group and its single members at the genomic level and, subsequently, ii) the lack of a proper identification of which specific HERV-W sequence is expressed in a given tissue/condition, currently prevents the assessment of the single HERV-W sequences expression in the different physio-pathological context, as well as their effective exploitation either as innovative diagnostic tools or therapeutic targets.

The present work was aimed to provide a comprehensive characterization of the HERV-W group in the human genome, setting the bases for a more rationale analysis of its members specific contribution to our transcriptome and to understand the potential effects exerted by HERV-W integrations and expressed products. In particular, a total of 213 HERV-W sequences were retrieved from human genome assembly GRCh37/hg19 and analyzed in terms of retroviral structure, genetic context of insertion, phylogeny and estimated time of integration; providing, to our knowledge, the most complete and exhaustive HERV-W dataset up to date, useful to clarify HERV-W role in pathologies with poorly understood etiology.

Moreover, to better understand the dynamics of diffusion of (H)ERV-W elements up to the human genome, we comparatively analyzed the group's presence and distribution in the genome sequences of 12 non-human primates, belonging to *Siimiformes* (*Catarrhines* and *Platyrrhines* parvorders), *Tarsiiformes* and *Prosimians*. Taken together the data obtained provide a complete analysis of the ERV-W group presence in the primate lineage, including

ERV-W loci orthologous to human integration as well as ERV-W species-specific insertions and events of LTR-LTR homologous recombination occurred along primates speciation, further detailing the group diffusion in primates genome. Interestingly, during the analysis in *Platyrrhini* species, we identified an ERV group named ERV1-1 that showed a high sequence identity to the ERV-W elements. The analysis of a set of the 130 most complete ERV1-1 sequences retrieved from Marmoset (*Callithrix jacchus*) and Squirrel Monkey (*Saimiri boliviensis*) genome assemblies allowed us to confirm the phylogenetic correlation of the ERV1-1 group to the ERV-W group, possibly suggesting the presence of a common ancestor of these Gammaretroviruses.

To gain more insights into the ERV1-1 group and its previously unreported relation with ERV-W group, the ERV1-1 sequences structure, context of insertion, phylogeny and estimated time of integration have been characterized. Such analysis provided, for the first time, a detailed description of the group, pointing out the various structural features shared with ERV-W elements, among which an unusual "so called" *pre gag* region.

Finally, considering that the ERV-W *pre gag* sequence is a stable component of the group members structure, for which neither origin nor function has been hypothesized yet, we analyzed in detail this region, and compared it with the ERV1-1 *pre gag*. Results showed that ERV-W and ERV1-1 *pre gag* regions share nucleotide identity in the 5'-portion, while ~40% of the ERV-W *pre gag* sequence appeared to be highly similar to a portion of the HERVIP10F ORF2, encoding for a Gammaretroviral *pol* protein Ribonuclease H domain. The same HERVIP10F-like *pre gag* region had been provided by the ERV-W sequences to Harlequin mosaic elements, possibly suggesting that a series of recombination events involved this ERV-W portion. Interestingly, in both ERV-W and ERV1-1 consensus sequences, the bioinformatics tool RetroTector detected the presence of a putative exon spanning the *pre gag* region. Considering that the predicted encoded peptides showed no homology with any known proteins, further studies are needed to assess the native role of the *pre gag* elements that have been maintained in these ERV groups and could be eventually present in other related Gammaretroviruses.

Chapter 1. INTRODUCTION

1.1 HUMAN ENDOGENOUS RETROVIRUSES (HERVs)

In the last 15 years, many efforts have been done to provide a complete sequence of the human genome, progressively revealing its unexpected, highly repetitive composition. Transposable elements (TEs) account in fact for >50% of our genetic material, while protein-coding regions constitute instead only the ~2% [1]. TEs can be broadly divided in two general classes based on their transposition intermediate (DNA or RNA). Human Endogenous Retroviruses (HERV) belong to class-I TEs, named also retrotransposons and characterized by a RNA intermediate that is reverse-transcribed into a double stranded DNA (dsDNA) and inserted at new genomic locations [2]. Beside HERVs, retrotransposons comprises also elements devoid of Long Terminal Repeats (LTRs), such as Long and Short Interspersed Nuclear Elements (LINEs and SINEs respectively), characterized by the presence of 3′ poly(A) repeats that appears critical for their retroposition [3].

HERVs are remnants of ancient retroviral infections acquired by the primates lineage in several waves, occurred mostly between 100 and 40 million years ago [4]. HERVs were once exogenous retroviruses whose infection affected not (only) the somatic cellular population, but, peculiarly, interested the germ line. Subsequently, the integration of the retrotranscribed proviral sequences within the germ line cells DNA made HERV sequences to become a stable part of our genome (Figure 1). Such process of endogenization and the further fixation in the host population allowed HERVs to be vertically transmitted to the offspring, being inherited in a Mendelian fashion up to constituting ~8% of the human genome [1].

In general, HERV sequences were formed by a traditional process of retroviral reverse transcription and integration, and show thus a classical proviral structure, with the *gag*, *pro*, *pol* and *env* genes flanked by two LTRs. Briefly, the *gag* gene encodes the structural components of matrix (MA), capsid (CA) and nucleocapsid (NC); the *pro-pol* genes determine the production of the three viral enzymes Protease, Reverse Transcriptase (RT) and Integrase (IN); and the *env* gene is responsible for encoding the envelope surface (SU) and trans-membrane (TM) elements. The 5′ and 3′ LTRs are formed during the retrotranscription process and are identical at time of integration. Due to their long time persistence into the host genome, HERV nucleotide sequences independently accumulated nucleotide substitutions, deletions and insertions, often leading to the loss of any coding capacity. In several cases, the homologous recombination between the two LTRs of a same provirus caused the complete elimination of the internal portion [5], a phenomenon reflected by the several thousand of solitary LTRs relics widespread in the human genome.

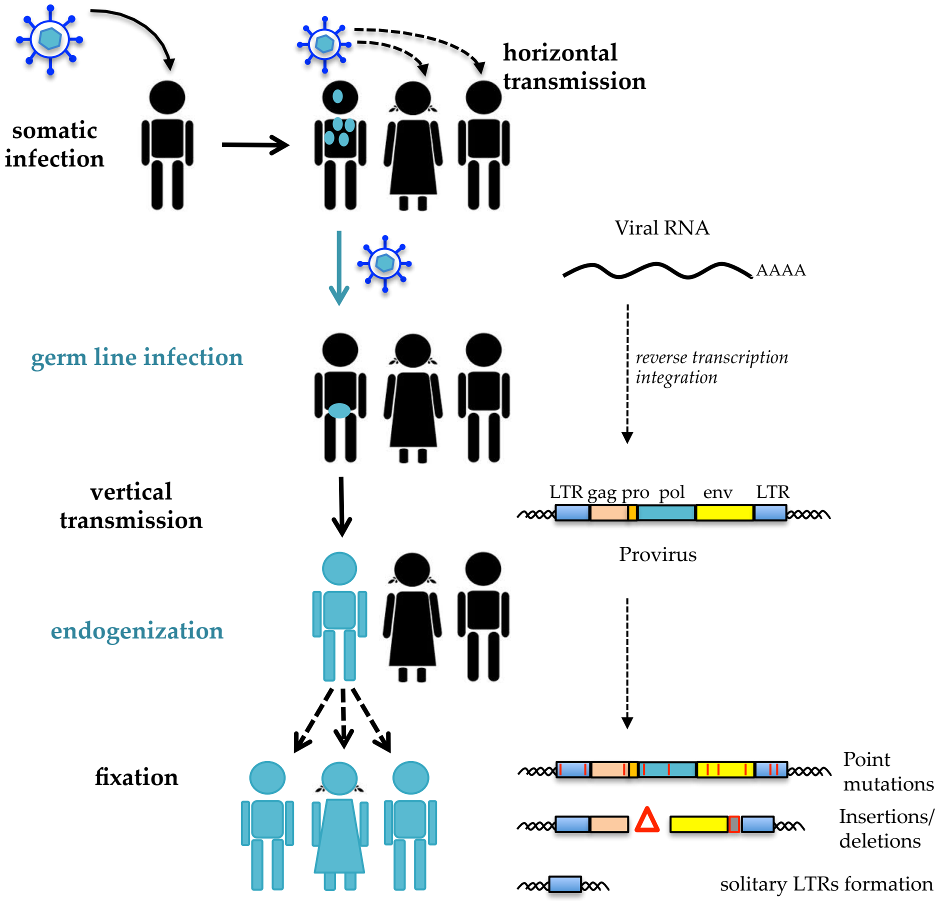


Figure 1. *Retroviruses endogenization and HERVs formation*
During replication, the retroviral RNA is reverse transcribed into a dsDNA provirus and stably integrated into the cellular genome. All known retroviruses target the somatic cells, showing a horizontal transmission from an infected individual to new hosts. The exogenous retroviruses that originated HERVs were indeed able to infect the germ line cells, which are transmitted to the offspring. In this way, also the integrated HERV sequences have been inherited in a Mendelian fashion, being endogenized in all the somatic and germ line cells. The subsequent vertical transmission of HERVs through the offspring determined their fixation into the human genome. During evolution, the majority of HERVs accumulated multiple mutations that generally compromised their coding capacity. In several cases, the homologous LTR-LTR recombination led to the elimination of the internal portion, leaving only a solitary LTR relic.

Despite their abundant presence, the general characterization of HERVs at the genomic level has been for a long time incomplete and sometimes controversial, due to the increasing amount of bioinformatics data and the concomitant absence of precise taxonomic rules for their classification [6]. Based on sequence similarity with the exogenous members, HERVs had been originally classified in three main classes: class I (Gammaretrovirus- and

Epsilonretrovirus-like), class II (Betaretrovirus-like) and class III (Spumaretrovirus-like). Each class encloses a variable number of groups, which have been named based on discordant criteria in the course of the years. In particular, HERV groups have been traditionally identified with a letter according to the type of human tRNA that binds the Primer Binding Site (PBS) to initiate the reverse transcription process [6]. For example, HERV-K elements are supposed to use a Lysine (K) tRNA, while HERV-W sequences PBS putatively recognizes a Tryptophan (W) tRNA. Some groups have also been occasionally named according to a neighbor gene (HERV-ADP) or a particular amino acid motif (HERV-FRD). Only very recently, the human genome assembly GRCh37/hg19 was analyzed with the use of the RetroTector program (ReTe) [7], leading to the recognition and first global classification of > 3000 HERV insertions [8] (Table 1). A multi-step classification approach based on similarity image analysis, *pol* gene phylogeny and taxonomic features allowed to characterize 39 "canonical" well defined groups of HERVs, and 31 additional "non canonical" clades describing several degrees of mosaicism, especially due to recombination and secondary integration events [8]. This comprehensive classification provided a remarkable background for the exhaustive characterization of single HERV groups at the genomic level, which still remains a major genetic and bioinformatics goal [9].

Table 1. *General HERV identification and preliminary classification in GRCh37/hg19 by ReTe*

HERV genus	RV genus-like	Type species	Sequences	Clades	HERV supergroups
Class I	GammaRV EpsilonRV	MLV, FELV, WDSV	2341	27 C 25 NC	HERV-T, HERV-ERI, HERV-W9, HERV-IPADP, HERV-HF, HERV-FRD-like, HEPSI, HUERSP
Class II	BetaRV	MMTV, MPMV, JSRV	598	10 C 0 NC	HERV-K (HMLs 1-10)
Class III	SpumaRV	SFV	216	2 C 5 NC	HSERVIII (HERV-S, HERV-L)
Uncertain	Errantivirus	Gipsy RV	2	0 C 1 NC	-
Unclassifiable	-	-	16	-	-

Adapted from Vargiu et al. 2016.
RV= retrovirus, MLV= Murine Leukemia Virus, FELV = Feline Leukemia Virus, WDSV= Walleye Dermal Sarcoma Virus, MMTV= Mouse Mammary Tumor Virus, MPMV= Mason-Pfizer Monkey Virus, JSRV= Jaagsiekte Sheep Retrovirus, SFV= Simian Foamy Virus, C= canonical, NC= non canonical.

Differently from the genomic characterization, which is still ongoing for most of the HERV groups, a large amount of studies - mainly based on microarrays, hybridization assays or RT-PCR approaches - assessed the various HERV groups general expression in a number of human healthy tissues and cell lines [10–17], revealing that HERVs sequences are stable components of our transcriptome. In particular, the presence of a differential expression among the diverse human tissues, and especially between healthy vs pathological samples, acted as a driving force for the search of a HERVs role in several human disorders, primarily cancer, inflammation and autoimmune and infectious diseases. Despite the great number of attempts, until now, the association of HERV expression to the diverse pathological conditions often ended in the field of "rumor-virology" [18], and no unequivocal cause-effect relationship has been established so far in any human disease [18–20]. Such current lack primarily depends on the absence of a proper HERV single groups characterization at the genomic side, essential to understand which precise HERV sequence is really expressed in a given condition [21] and if this expression is beneficial, maleficent or just functionally linked to a specific condition, and hence potentially exploitable as diagnostic biomarker or therapeutic target. Moreover, it is important to consider that many of the diseases for which an HERV role has been proposed are chronic conditions with poorly understood etiology, with several other factors (genetic or environmental) that could be required and concur to a causal association [18]. All these aspects are to be addressed into the wide panorama of disparate HERVs expression data, including a large amount of discordant findings and very few studies attempting to link the various observations in the same experimental groups and conditions [18].

1.2 THE HERV-W GROUP

Among HERVs, the HERV-W group is one of the most intensively investigated with regards to its expression and the possible physio-pathological effects on the host. Initially identified as possible causative agent for Multiple Sclerosis (MS) [22], the HERV-W group was found to be strongly expressed in placental tissues [23]. This observation led to the identification of a single member (ERVWE1, locus 7q21.2) still able to encode a functional Envelope (Env) protein, which has been coopted during evolution for an important function in placentation [24, 25]. If, on the one side, this HERV-W element and its physiological role have been described in great detail [26–31], on the other side, the general expression of the HERV-W group has been broadly investigated in a variety of tissues, above all to find a correlation to a wide range of diseases. However, the observed expression profiles have not been linked to any specific HERV-W sequence in the great majority of cases, and no definitive association with any human pathology has been conclusively demonstrated so far.

Commonly to the other HERV groups, the HERV-W sequences acquisition by the primates germ line was initially due to a traditional process of retroviral infection (Figure 2). In general, it is not clear whether the exogenous retroviruses originating HERVs had germ line cells as specific target or infected such population by chance [32]. In any case, after the entry into germ line cells, the viral RNA genome was reverse transcribed into proviral dsDNA, flanked by identical LTRs and competent for the subsequent integration into the host cell genome. Several of these repeated replication events determined the initial diffusion of HERV-W insertions among human chromosomes, with new proviruses formation possibly occurring also in the absence of an extracellular infectious phase [35], by proviruses expression and integration of the reverse transcribed transcripts [32] (Figure 2). Worth of note, differently from other known HERV groups, the HERV-W spreading was also in larger part sustained by L1-mediated processed pseudogenes formation [33, 34] (Figure 2). L1, or LINE-1, are Long Interspersed Elements belonging to the non-LTR retrotransposons and encode for a protein with RT and endonuclease activity [35]. They are able to copy and paste their own RNA into new genomic sites, but also to mobilize other non-LTR retrotransposons (Alu and SVA) as well as the expressed HERV-W elements. In this way, RNA transcripts originated from preexisting HERV-W proviral insertions have been reverse transcribed and integrated in a different chromosomal position by the L1 machinery. These elements, named processed pseudogenes, are characterized by some specific sequence signatures, differing from the traditional proviruses and structurally resembling a viral mRNA: i) truncated 5′ and 3′ LTRs, showing a R-U5 and U3-R structure instead of the traditional U3-R-U5 one, respectively; ii) a poly(A) tail of variable length, and iii) a TT/AAAA insertion motif and a 5–15 nucleotides target site duplication [21, 33]. It is worth to note that the L1-mediated processed pseudogenes formation was not a minor event in HERV-W spreading: these sequences account, in fact, for the majority of the so far identified HERV-W [21, 33, 34]. Of further note, the molecular model of L1-recognition and retrotransposition of HERV-W sequences, as well as the specificity determinants that limited this process to the HERV-W group transcripts remain to be clarified.

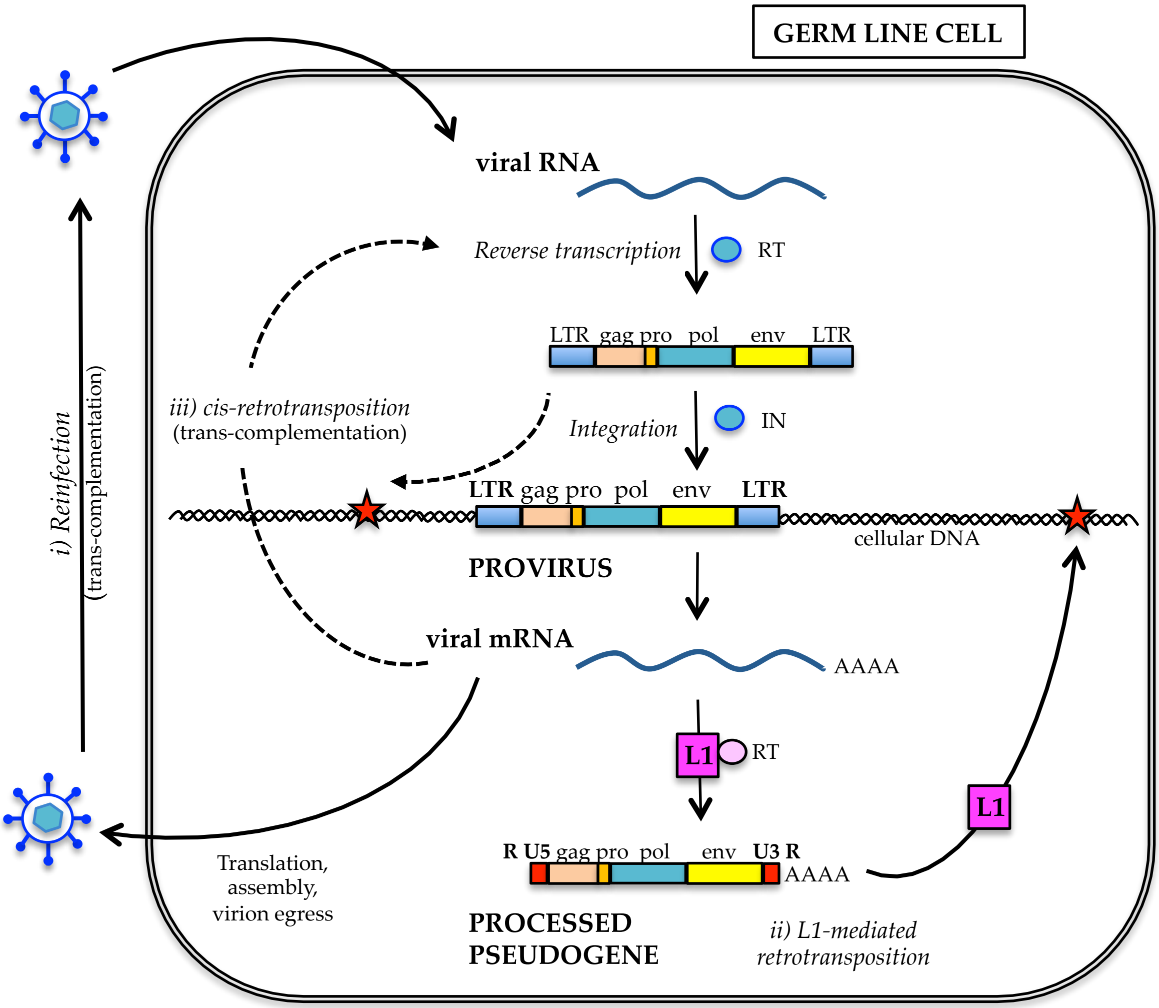


Figure 2. *Overview of the HERV-W integration pathways within the human germ line cells genome.*
The initial acquisition of HERV-W sequences has been due to a traditional retroviral infection: the genomic RNA was reverse transcribed and integrated into the host cell genome, forming HERV-W proviruses. The integrated proviruses transcription led to the production of viral mRNAs, which could generate new HERV-W insertions (red stars) through i) Reinfection: proviral mRNAs have been translated and the deriving proteins assembled into a mature viral particle, that after egress could be able to re-infect the cell, providing new HERV-W integrations; ii) L1- mediated retrotransposition: copy-and paste mechanism in which viral mRNAs have been reverse transcribed and mobilized by L1 machinery, being inserted into a new genomic position as processed pseudogene with a mRNA-resembling structure; iii) *Cis*-retrotransposition: mRNAs could also have been used as templates for a new intracellular reverse transcription-integration process, leading to new proviruses acquisition in the absence of an extracellular phase. Due to the accumulation of mutation within the HERV-W proviruses coding structures over time, these last two mechanisms could have required the presence of proteins provided in *trans* by a helper virus.

With regards to the HERV-W group characterization at the genomic level, until very recently the main references for the field were three independent studies performed a number of years ago on isolated human chromosomes [36] or incomplete draft versions of the human genome [33, 34]. Despite these studies represent milestones in the analysis of HERV-W group, the use of different detecting methodologies led to discordant results at the time, currently difficult to retrieve and especially to correlate with modern data, such as the abundantly available RNA-seq expression profiles.

1.3. HERV-W GROUP PHYSIOLOGICAL EXPRESSION

1.3.1 Placental expression and Syncytin-1 protein cooption

First evidences suggesting the presence of retroviral particles with RT activity in MS patients samples [37, 38] (see below) led to the first description of the so called MS Retrovirus (MSRV). Subsequent Southern blot analysis using MSRV-derived probes allowed to detect a previously undescribed HERV multicopy family [22], formally named HERV-W group [23]. The molecular characterization of the group coding capacity interestingly revealed a strong expression restricted to placenta [24] (apart from minor expression in testis [25]), showing the presence of a complete ORF encoding for two 4 and 8 kb major transcripts [24, 25] and producing a 538 amino acids functional Env protein suggested to have a role in pregnancy-related physiological functions [23]. The identified Env was mapped to a HERV-W locus on chromosome 7q21.2 (ERVWE1) [25], which harbored a 5'LTR functional promoter exhibiting several binding sites for transcriptional regulators involved in the control of proliferation and differentiation [26, 39]. This HERV-W Env was expressed *in vitro* in a panel of different species cell lines, resulting to be able to interact with the type D mammalian retrovirus receptor (hASCT2, human sodium-dependent neutral amino acid transporter type 2) strongly inducing syncytia formation [24, 25], and was therefore named Syncytin-1 [25]. The evidence that syncytia formation could be specifically impaired with both an anti-Syncytin-1 antibody [24, 25] and anti-sense inhibition of Syncytin-1 expression [40] confirmed a central role for this HERV-W Env in the homo- and hetero-typic fusogenicity [24, 25]. Despite the cell-cell fusion induction by Syncytin-1 primary depends on the interaction with hASCT2 receptor [24], further studies showed that this Env can efficiently bind both hASCT1 and hASCT2 receptors [41]. Moreover, even the highly divergent mouse orthologous transporters, mASCT1 and mASCT2, could be recognized after the elimination of their N-linked glycosylation sites [41], suggesting a lower restriction of receptor utilization than the one other retroviruses Env proteins, likely due to the strong selective pressure throughout evolution [41]. Syncytin-1 placental expression was specifically confirmed in the villous [25] and extravillous trophoblasts [42], and its strong fusogenic activity was associated with the formation of villous syncytiotrophoblast, which is the primary site for trophic exchanges and many other placental functions fundamental for fetal growth and development [24, 25]. Beside this central fusogenic role, Syncytin-1 was shown to be directly involved also in the upstream primary cytotrophoblast differentiation and proliferation, essential for the syncytiotrophoblast homeostasis [40, 43]. Cyclic AMP (cAMP), which regulates cAMP-dependent protein kinases in trophoblast fusion and differentiation [44], was in fact also able to control Syncytin-1 expression [40]. This regulation depends on the Syncytin-1 promoter, constituted of a bipartite element formed by i) the ERVWE1 5′LTR, holding the basal placental promoter activity and cAMP-responsive elements, adjacent to ii) a placenta-restricted cellular enhancer, located within a MaLR solo LTR, that function as an URE (Upstream Regulatory Element), conferring high levels of tissue-specific expression [29]. Syncytin-1 siRNA knockdown in BeWo cultures was shown to reduce cell growth and proliferation, and this seems to occur through the cell cycle arrest in G1 phase [43]. Contrarily, the ectopic overexpression of Syncytin-1 appeared to positively modulate trophoblast cells proliferation, confirming the critical role of this protein in promoting the G1/S transition during syncytiotrophoblast formation, and emphasizing a subtle balance of fusogenic and non-fusogenic functions in the co-regulation of the cytotrophoblast pool [43]. Finally, in addition to cell-cycle regulation, Syncytin-1 seems to play a role also in the control of trophoblast survival, since expression knockdown in BeWo cells resulted in triggering apoptosis by specific activation of the apoptosis inducing factor (AIF) mediated cell death pathway [45].

In addition to these important functions in placental morphogenesis and homeostasis, Syncytin-1 was also hypothesized to have a role in the maternal immunotolerance to the fetus [24, 25, 42] through its transmembrane immunosuppressive domain [25], as previously demonstrated *in vivo* for the Env proteins of a murine [46] and a primate [47] retrovirus. While subsequent studies reported the absence of such activity in a mice model, suggesting a genetic disjunction between fusogenic and immunosuppressive functions (at least in mice) [48], Syncytin-1 in human blood was able to effectively inhibit the production of Th1 cytokines, known to be important modulators of several immunological functions. This suggests a Syncytin-1 possible role in mediating the shift from Th1 to Th2 cytokines observed during pregnancy, thus likely contributing to the immunomodulation of the maternal system [49].

Overall, these observations revealed a pivotal role of Syncytin-1 expression in placental morphogenesis, being directly involved in many aspects of the syncytiotrophoblast formation and homeostasis through a well-evolved balance of both fusogenic and non-fusogenic functions. Differently from other HERV-W members, due to its relevance for human physiology, the Syncytin-1 HERV-W locus as well as its Env gene and protein have been structurally characterized in great detail in humans and in hominoids and great apes primates [26–31].

1.3.2. HERV-W expression in other healthy tissues

Syncytin-1 locus constitutes a very remarkable exception of a HERV retaining a residual protein-coding capacity, while the great majority of HERV sequences apparently accumulated mutations determining the loss of the ability to produce proteins. For this reason, HERVs have been often disregarded in large scale expression studies and, subsequently, not exhaustively characterized in terms of expression, regulation and functional significance [50]. A number of studies, however, investigated their expression across human tissues and cells, revealing that HERVs are stable components of our transcriptome. Overall, results showed that HERV groups are characterized by differential global expression profiles, which could be tissue/cell type-specific and vary depending on the tissue state changes (e.g. differentiation, pathogenesis) and environmental and individual conditions.

As stated above, HERV-W group shows a strong expression in normal placenta [23–25], and a significant transcriptional activity in testis [25]. Beside these results, the analysis of HERV-W transcription in healthy tissues has been performed by several other independent studies, mainly by the use of RT-PCR amplifying either *gag*, *pol* or *env* genes with primers that, in the great majority of cases, were specifically designed on placental Syncytin-1 ERVWE1 locus. Other few studies searched Expressed Sequence Tags (ESTs) databases using Syncytin-1 provirus sequence as a query, or analyzed the group expression through *pol* probes hybridization. In this way, a general HERV-W expression has been detected in various human cell lines and healthy tissues, mostly lacking, however, any information about the locus of transcripts origin (Table 2).

Table 2. *General HERV-W group expression in non-placental healthy tissues*

Tissue	Method	Ref.	Possible biases of HERV-W members underrepresentation[a]
Blood	Search of Syncytin-1 query in EST data	[11]	Low total HERV EST counts, undetection of HERV-Ws divergent from Syncytin-1, no information on LTR activity, n°of cDNA/EST libraries great variability across tissues, under-representation of poorly expressed genes in small libraries (1)[b]
Brain	Search of Syncytin-1 query in EST data	[11]	(1)
	RT-PCR (*gag+*, *pol+*, *env +)*	[51]	Primers specific for single expressed sequences (placental Syncytin-1 (*gag:* AF072500, *env:* AF072506), MSRV clones (*pol:* AF009668)) could undetect divergent HERV-Ws, no information on full-length HERVs expression and LTR activity, samples amount is poorly representative (2)
Brain (cortex and pons)	*env* real time qRT-PCR	[52]	Primers specific for placental Syncytin-1 (NM_014590.3) can undetect *env* defective or highly divergent HERV-Ws, no information on full-length HERVs expression and LTR activity, samples amount is poorly representative (3)
Breast	Search of Syncytin-1 query in EST data	[11]	(1)
	env real time qRT-PCR	[52]	(3)
Colon	*env* real time qRT-PCR	[52]	(3)
Heart	RT-PCR (*gag-*, *pol-*, *env +)*	[51]	(2)
Endometrium	GammaHERV and HERV-W *pol*-based probe and probe-less real time qPCRs	[53] [54]	Undetection of transcripts defective or highly divergent for *pol* gene, no information about full-length sequences expression and LTR activity, samples amount is poorly representative (4)
Kidney	*pol*-expression arrays hybridization	[55]	Cross-amplification/hybridization of related HERV groups; undetection of transcripts defective for *pol* gene, no information about full-length sequences expression and LTR activity, no quantitative information, samples amount is poorly representative (5)
	RT-PCR (*gag-*, *pol+*, *env +)*	[51]	(2)
Liver	*pol*-expression arrays hybridization	[55]	(5)
	RT-PCR (*gag-*, *pol+*, *env +)*	[51]	(2)
	env real time qRT-PCR	[52]	(3)
Liver-spleen (fetal)	Search of Syncytin-1 query in EST data	[11]	(1)
Lung	RT-PCR (*gag-*, *pol+*, *env +)*	[51]	(2)
Ovary	Search of Syncytin-1 query in EST data	[11]	(1)
	GammaHERV and HERV-W *pol*-based probe and probe-less real time qPCRs	[53] [54]	(4)
PBMC	*pol* RT-PCR and *env* real time PCR	[17]	Low sensitivity and cross-amplification of related HERV groups by RT-PCR degenerate primers, real time PCR primers specific for placental Syncytin-1 (NM_014590.3) can undetect divergent HERV-Ws, undetection of transcripts defective for *pol/ env* genes, no information on full-length sequences expression and LTR activity, incomplete characterization of individuals health status
Prostate	RT-PCR (*gag-*, *pol+*, *env +)*	[51]	(2)
Skel. muscle	RT-PCR (*gag-*, *pol+*, *env +)*	[51]	(2)
Spleen	RT-PCR (*gag+*, *pol+*, *env +)*	[51]	(2)
Stomach	*env* real time qRT-PCR	[52]	(3)
Testis	RT-PCR (*gag+*, *pol+*, *env +)*	[51]	(2)
Thymus	RT-PCR (*gag-*, *pol+*, *env +)*	[51]	(2)
Uterus	RT-PCR (*gag-*, *pol-*, *env +)*	[51]	(2)
	env real time qRT-PCR	[52]	(3)

[a] Methodological biases that potentially affected the effective and specific detection and characterization of the expressed HERV-W sequences. After the first mention, biases with multiple citations are reported with the listing number.

In particular, the global HERV-W expression was reported by Stauffer et al. in blood, brain, breast liver/spleen, ovary and placenta, even if the total EST counts were relatively low, and a subsequent analysis confirmed such results for placenta and breast tissues only [11]. The presence of a physiological HERV-W env transcription in healthy brain and breast was detected also by Kim et al. [52]. Yi et al. investigated the HERV-W gag, pol and env genes expression within 12 tissues (brain, prostate, testis, heart, kidney, liver, lung, placenta, skeletal muscle, spleen, thymus and uterus), reporting env transcripts in all the analyzed samples and some tissue-specific expression for gag (brain, testis, placenta and spleen) and pol (all tissues except for heart and uterus) [51]. HERV-W RNA expression was reported also in normal endometrium and ovary [53, 54], colon, liver, stomach and uterus [52]. The HERV-W group (together with HERV-H and HERV-K) was found to be transcriptionally active in peripheral blood mononuclear cells (PBMC) since early childhood, with a significant increase in subjects >40 years old [17]. High resolution melting temperature analysis [56] assessed the occurrence of systematic variations in the expression of HERV-W gag sequences in primary fibroblasts, depending on both tissues and individuals considered [57].

In summary, a global HERV-W transcriptional activity in healthy conditions was reported by at least one study in brain, breast, skeletal muscle, spleen, lungs, digestive trait (stomach, liver, colon), genitourinary apparatus (ovary, endometrium, uterus, prostate, testis and kidneys) and cardiovascular system (heart, whole blood, PBMC). Noteworthy, all these reports assessed the HERV-W group generic expression, i.e. without connecting the observed transcripts to a specific locus, and could be possibly biased by the use of Syncytin-1 provirus/MSRV cDNA clones as a query and for the design of primers and probes. This can lead, in fact, to the lack of detection of HERV-W expressed loci with divergent nucleotide sequence, or defective for the single genes analyzed. Moreover, in the majority of cases, no information about the full-length HERV-W sequences expression and the LTR residual activity are available, and the samples amount is poorly representative and sometimes incompletely characterized regarding the individuals health status.

Differently, an attempt to connect HERV transcriptome to specific loci of origin was performed by Pérot et al. through a dedicated microarray designed on a collection of > 5500 HERVs (both proviruses and solitary LTRs) that can reasonably be allocated to unique genomic loci [15] (Table 3). Authors suggested that HERVs transcriptome depends on tissue tropism and is sensitive to tissue state changes, including in the study a small set of paired normal vs tumoral tissues [15]. Based on their results, the HERV-W group showed an expected, predominant expression in placenta and testis, attributable to Syncytin-1 locus activity. In addition, 5 other specific HERV-W loci (1 provirus, 1 processed pseudogene and 3 solitary LTRs) were also transcribed in the same two tissues, showing in two cases a concomitant LTR promoter activity [15]. Despite the tissues considered by Pérot et al. were limited (colon, lung, breast, ovary, prostate, testis, uterus and placenta) and all the expressed HERVs were co-localized within human genes that could influence their transcription, the analysis is a remarkable effort to match HERV transcriptome to its specific genomic contributors, taking into account relevant aspects such as promoter activity and tissue specificity.

Table 3. *Specific HERV-W loci for which expression in healthy tissues has been reported*

Locus	**Chr:start-end (strand)**[a]	**Type**	**Genomic context**[b]	**Tissue**	**Method**	**Ref.**
2q22.1	2:139030735-139031481 (-)	Solo LTR	*LTR8 (+)*	Testis	Microarray	[15]
2q24.3	2:165514421-165516121 (-)	Pseudogene	COBLL1 (-), *TCONS_00004484 (-)*	Placenta	Microarray	[15]
5q12.1*	5:59954322-59962280 (+)	Provirus	DEPDC1B (-)	Placenta	Microarray	[15]
7q21.2*	**7:92097313:92107506** (-)	Provirus	-	Placenta Testis	Northern Blot	[23] [25]
15q21.2	15:51552784-51553570 (+)	Solo LTR	CYP19A1 (-)	Placenta	Microarray	[15]
Xq21.33	X:93824238-93824702 (-)	Solo LTR	MER4A (-)	Placenta	Microarray	[15]

[a] Chromosomal positions are referred to genome assembly GRCh37/hg19. Syncytin-1 locus is highlighted in bold.
[b] Localization of HERV-W loci within a human gene (italic names correspond to non-coding genes).
For sequences marked with an * a LTR promoter activity has been also reported.

1.4 HERV-W GROUP EXPRESSION IN THE PATHOLOGICAL CONTEXT

1.4.1 Syncytin-1 expression in placental pathologies

Consistently with its proven role in human placentation, an abnormal Syncytin-1 expression has been observed in various pathological conditions affecting placental and maternal-fetal physiology, i.e. Pre-Eclampsia (PE); Hemolysis Elevated Liver enzymes and Low Platelet count (HELLP) syndrome; Trisomy 21; IntraUterine Growth Restriction (IUGR) and endometriosis. The main findings in these pathological contexts are summarized below.

PE is a multisystem condition affecting ~5 % of pregnant women [58], clinically characterized by hypertension, proteinuria and hypoxia, and associated with adverse perinatal outcome and preterm birth. A significant reduction in trophoblast cells fusion isolated from PE placentas was reported [59] and, in line with this observation, placentas of women affected by PE showed a marked decrease in Syncytin-1 expression [59–63]. Such reduction seems to correlate with PE severity [59] and to depend on Syncytin-1 promoter hypermethylation [63], leading to consequent decrease in cytotrophoblast differentiation [39].

Similar results were found in HELLP syndrome [60, 61], another pathology exclusive of pregnancy. Considering that experimental hypoxia reduces the Syncytin-1 expression by 80% in BeWo cells *in vitro* and in isolated placental cotyledons *ex vivo* [64], it has been suggested that reduction in Syncytin-1 expression might arise secondarily to the failure of arterial transformation by trophoblast and the poor placental perfusion, as common in PE [65].

Another similar observation was done in trophoblast cells from placentas bearing trisomy 21 fetus, that still are able to aggregate but fuse poorly or late in culture [66–68] and in which the levels of superoxide dismutase, encoded on chromosome 21, are increased [69]. When this antioxidant enzyme was hyper-expressed in normal cytotrophoblasts, impairment in syncytiotrophoblast formation as well as abnormal cell fusion and Syncytin-1 transcription decrease were observed, further suggesting an influence of oxidative states on Syncytin-1 production [66, 69]. Since it is known that hypoxia can activate the caspase apoptotic pathway, the hypoxic environments common to many placental diseases could possibly lead to trophoblast cell death via both this mechanism and the above mentioned AIF pathway [43], specifically triggered by Syncytin-1 decreased expression [43, 58].

IUGR is another important cause of perinatal morbidity and mortality for both mother and fetus related to hypoxia and abnormal trophoblast development. In line with this, IUGR placentas showed significantly lower Syncytin-1 RNA and protein levels with respect to control placentas [62, 70], however still sufficient to mediate cells fusion [70].

Finally, some data reported the presence of high HERV-W expression in endometriotic tissues, even though no great differences were found with respect to control tissues [53, 54], and the Syncytin-1 upregulation, through the hypometilation of its promoter region, has been proposed to be involved in the development of endometriotic lesions [71].

Overall, these findings confirmed a pivotal role of Syncytin-1 expression in placental physiology, showing how this Env protein deregulation could contribute to the close connection between syncytial structure and maternal systemic disorders [58].

1.4.2 Tumorigenesis and cancer progression

Tumorigenesis is a complex multistep process likely involving both inherited and environmental factors, and is another highly investigated field for the possible association of HERVs expression to human pathogenesis. Of course, this hypothetical link has been greatly sustained by the well-described transforming nature of various animal exogenous retroviruses, originally designated as "RNA tumor viruses". However, in contrast to exogenous retroviruses, the high copy number and repetitive nature of HERVs in the host genome may trigger additional tumorigenesis mechanisms that do not require the production of infectious viral particles, as summarized in Figure 3 panels. In particular, similarly to other TEs, HERVs mobilization and integration could be responsible for insertional mutagenesis events (panel a), which could disrupt or deregulate host genes (e.g. oncosuppressors, transcriptional regulators). The presence of repetitive elements could also trigger chromosomal rearrangements by non-allelic homologous recombination (panel b). HERV sequences transcriptional de-repression, possibly prompted by the altered epigenetic environment commonly associated to cancer tissues, can activate LTR promoters and subsequently lead to uncontrolled activation of downstream cellular genes (e.g. oncogenes, transcription factors) (panel c). Also in the absence of protein production, HERVs transcription could determine the accumulation of incomplete replication intermediates, able to activate innate immunity pathways or deregulate non-coding RNA regulatory networks (panel d). Finally, if a protein (or a portion of it) is produced, its native functional activities (e.g. fusogenic and/or immunosuppressive functions) as well as its eventual ability to interact with cellular proteins may contribute to tumor development (panel e).

Regarding the latter case, direct mechanisms of contribution to tumor development have been proposed for Rec and Np9, two accessory proteins produced by HERV-K(HML2) elements through alternative splicing of the *env* gene. These proteins are often specifically reported in transformed cells, and Rec is able to interact with the cellular transcription factor PLZF and induce tumors in mice if overexpressed [72–74].

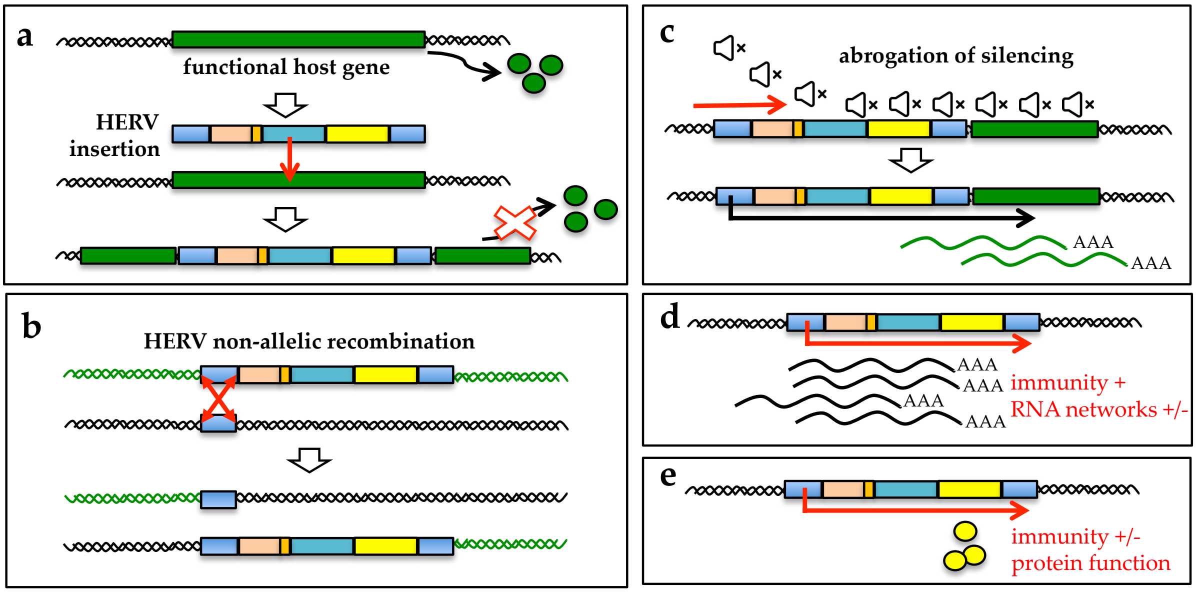


Figure 3. Potential mechanisms of HERV-mediated transformation in tumorigenesis.
a) Insertional mutagenesis can disrupt/deregulate host genes;
b) non-allelic homologous recombination can induce chromosomal rearrangements;
c) transcriptional silencing abrogation can trigger LTR promoter activity;
d) accumulation of replication intermediates can evoke immunity and/or deregulate RNA networking;
e) protein production can evoke immunity and/or provide oncogenic functions.

Remarkably, despite several works reported a general increase or a *de novo* appearance of HERV expression in tumors with respect to the correspondent healthy tissues (Table 4), it is yet to be understood whether such expression is the cause or just the consequence of transformation processes. In fact, while in adult healthy cells HERV expression is generally silenced by epigenetic mechanisms, the abnormal hypomethylation of CpG dinucleotides is commonly observed in many tumoral tissues, generally associated with increased levels of expression. Due to this dysregulation, the HERVs hyper-expression observed in cancer could likely be, in part, an indirect product of the altered transcriptional and epigenetic environment of cancer itself, instead of a determinant of disease onset. In line with this thesis, a study investigating the methylation state of L1 and HERV-W retrotransposons in human ovarian carcinomas reported the consistent reduction of promoter CpG methylation, correspondent to an increase in the relative L1 and HERV-W expression [75]. Such upregulation, involving both L1 and HERV-W expression, could of course contribute to tumor progression also by the *de novo* mobilization of the abundant HERV-W transcripts in cancer cells, as a possible source of processed pseudogenes mutagenic insertions likely sustaining the oncogenic process.

***Table 4**. General HERV-W group expression in tumoral tissues*

Tumoral Tissue	Ref.	Method[b]	Basal status[a]	Possible biases of HERV-W members underrepresentation[a]
B cells	[51]*	RT-PCR (*gag-, pol-, env+)*	[17]°	Primers specific for single expressed sequences (placental Syncytin-1 (*gag:* AF072500, *env:* AF072506), MSRV clones (*pol:* AF009668)) could undetect divergent HERV-Ws, no information on full-length HERVs expression and LTR activity, samples amount is poorly representative (2)
Bladder	[51]*	RT-PCR (*gag-, pol+, env+)*	-	(2)
Breast	[11]	Search of Syncytin-1 in EST data	[11, 52]	Low total HERV EST counts, undetection of HERV-Ws divergent from Syncytin-1, no information on LTR activity, n°of cDNA/EST libraries great variability across tissues, under-representation of poorly expressed genes in small libraries (1)
	[76]*	RT-PCR, real time qRT-PCR,		Specific detection of a Syncytin-1 *env* portion only, undetection of transcripts divergent/defective for *env*, no information on full-length sequences expression and LTR activity
	[52]	*env* real time qRT-PCR		Primers specific for placental Syncytin-1 (NM_014590.3) can undetect *env* defective or highly divergent HERV-Ws, no information on full-length HERVs expression and LTR activity, samples amount is poorly representative (3)
	[51]*	RT-PCR (*gag-, pol+, env +)*		(2)
Brain	[51]*	RT-PCR (*gag-, pol+, env +)*	[11, 51]	(2)
Colon	[11]	Search of Syncytin-1 in EST data	[52]	(1)
	[52]	*env* real time qRT-PCR		(3)
	[51]*	RT-PCR (*gag-, pol+, env +)*		(2)
	[77]*	qPCR		Specific detection of a Syncytin-1 *env* portion only, undetection of transcripts divergent/defective for *env*, no information on full-length sequences expression and LTR activity
Endometrium	[78]	qPCR, RT-PCR, NB, WB	[53, 54, 78]	Specific detection of a small portion of Syncytin-1 *env* only, samples amount is poorly representative, expression values are highly heterogeneous
Esophagus	[51]*	RT-PCR (*gag-, pol+, env +)*	-	(2)
Histiocyte	[51]*	RT-PCR (*gag-, pol+, env +)*	-	(2)
Kidney	[11]	Search of Syncytin-1 in EST data	[51, 55]	(1)
	[51]*	RT-PCR (*gag-, pol+, env +)*		(2)
Liver	[52]	*env* real time qRT-PCR	[51, 52, 79]	(3)
	[51]*	RT-PCR (*gag-, pol+, env +)*		(2)
Lung	[51]*	RT-PCR (*gag-, pol+, env +)*	[51]	(2)
Neuroblasts	[80, 81]*	*pol* real time qPCRs	-	Undetection of transcripts defective or highly divergent for *pol* gene, no information about full-length sequences expression and LTR activity, samples amount is poorly representative (4)
Ovary	[75]	Real time qRT-PCR	[53, 75]	Primers designed on Syncytin-1 locus (AC000064) can undetect divergent HERV-Ws, samples amount is poorly representative
	[53]	*pol* real time qPCRs		(4)
	[51]*	RT-PCR (*gag-, pol+, env +)*		(2)
Pancreas	[51]*	RT-PCR (*gag-, pol+, env +)*	-	(2)
Placenta	[11]	Search of Syncytin-1 in EST data	[23–25]	(1)
Prostate	[51]*	RT-PCR (*gag-, pol-, env +)*	[51]	(2)
Skin	[51]*	RT-PCR (*gag-, pol-, env +)*	-	(2)
Stomach	[52]	*env* real time qRT-PCR	[52]	(3)
	[51]*	RT-PCR (*gag-, pol+, env +)*		(2)
T-cells	[51]*	RT-PCR (*gag-, pol+, env +)*	[17]°	(2)
Uterus	[52]	*env* real time qRT-PCR	[51, 52]	(3)
	[51]*	RT-PCR (*gag-, pol+, env +)*		(2)

[a] Based on studies which reported the general group expression in the correspondent healthy tissues.

[b] NB= Northern Blot, WB= Western Blot.

[c] Methodological biases that potentially affected the effective and specific detection and characterization of the expressed HERV-W sequences. After the first mention, biases with multiple citations are reported as a number.
° data obtained in total PBMC. * data obtained in cancer cell lines

Interestingly, despite the hypomethylation affected the great majority of the analyzed elements, some L1 and HERV-W sequences remained hypermethlyated in the malignant samples [75], suggesting that, even if a great proportion of HERV-W elements could be transcribed as a consequence of the tumor environment, a portion of loci could have indeed a specific role, either beneficial or detrimental, in the disease progression. HERV-W elevated transcription in ovarian carcinomas has been reported also by Hu et al., but a similar expression level was observed also in healthy ovary control tissues, and the number of samples was too low to be statistically significant [53].

Some other studies reported HERV-W group expression in a number of tissues, showing however a different scenario. Stauffer et al. investigated HERV-W expression in placenta, breast, colon and kidney cancers; however, while no HERV-W basal expression was found for healthy colon and kidney, the HERV-W transcription levels in breast and placenta were even higher in healthy tissues than in tumor samples [11]. Paired tumoral vs healthy tissues were analyzed also by Kim et al., reporting no significant differences of HERV-W expression between tumor and normal adjacent tissues in breast, colon, liver, stomach and uterus [52].

While the previous analyses were performed on patients' tumoral samples, a number of other studies investigated the HERV-W transcriptional activity in tumors through their detectable expression in the correspondent cancer cell lines. Yi et al. interestingly assessed HERV-W group transcription in various tissues reporting, in some cases, a lack of correlation between the expression observed in normal cells with respect to the correspondent cancer cell line [51]. Bjerrgarden et al. showed that MCF-7 and MDA-MB-231 breast cancer cells express Syncytin-1 and hASCT-2 receptor on the external surface, being able to fuse with endothelial cells presenting only the hASCT-2 receptor, hypothesizing a Syncytin-1-mediated pathological fusogenic activity for breast cancers, further confirmed by the lack of fusion when Syncytin-1 expression was downregulated [76]. When considering brain tumors, HERV-W RNA levels were increased in 3 neuroblastoma cell lines (SH-SY5Y, SK-N-DZ and SK-N-AS), showing a selective upregulation during hypoxia recovery and after the treatment with demethylating agents [80, 81]. In another study, SH-SY5Y and another neuroblastoma line transfected with HERV-W *env* showed the overexpression of SK3 (small conductance Ca2+- activated K+ channel protein 3), an ion channel relevant for neuronal excitotoxicity and linked to various nervous system diseases [82]. Such upregulation was proposed to depend on the activation of SK3 promoter cAMP responsive element (CRE) sites, in line with the HERV-W Env mediated increased phosphorylation of the relative activating transcription factor CREB (CRE-binding protein) [82]. Finally, after the reported Syncytin-1 expression in colon adenocarcinoma patients, Dìaz-Caballo et al. reported the HERV-W expression in HCT8 colon carcinoma cells, proposing its link to the induction of a chemotherapy-refractory status [77].

As previously reported for HERV-W physiological expression, while many reports assessed the general HERV-W altered transcription in different tumor tissues, a very few studies attempted to connect it to specific HERV-W loci of origin (Table 5).

***Table 5**. Specific HERV-W loci reported as hyperexpressed in tumoral tissues*

Locus	**Chr:start-end (strand)**[a]	**Type**[b]	**Genomic context**[c]	**Tissue**[d]	**LTR**[e]	**Method**	**Ref.**
1q31.2	1:192855545-192856320 (-)	LTR	MER21C (-)	Testis	-	MA[f]	[15]
2p24.2	2:17520208-17527981 (+)	PV	L3 (-)	Testis	Pro°	MA, qRT-PCR	[15, 83]
2p12	2:76098816-76106624 (+)	PV	-	Testis	Pro	MA	[15]
3p12.3	3:74921984-74927237 (-)	PG	-	Prostate	-	MA	[15]
3q28	3:191376573-191383381 (+)	PG	-	Testis	-	MA	[15]
4p13	4:42287455-42294913 (-)	PV	TCONS_00007753 (-)	Testis	Pro°	MA, qRT-PCR	[15, 83]
4q26	4:114965536-114972972 (+)	PG	-	Testis	-	MA	[15]
5p13.3	5:31109366-31109859 (-)	LTR	-	Ovary	-	MA	[15]
6q21	6:106676012-106683689 (+)	PG	ATG5 (-)	Skin T cells	-	MA, qRT-PCR	[84]
7q21.2	**7:92097313:92107506 (-)**	**PV**	-	Testis*	Pro°	MA, qRT-PCR	[83]
				Bladder	Pro	qRT-PCR	[85]
				Skin T cells	-	MA, qRT-PCR	[84]
7q21.3	7:95987661-95988433 (-)	LTR	Alu Sx (-)	Testis	-	MA	[15]
7q31.1b	7:114019143-114026368 (-)	PG	*FOXP2 (+)*	Testis	-	MA	[15]
7q36.3	7:155177752-155178503 (-)	LTR	BC150495 (+)	Testis	PA	MA	[15]
8q24.13	8:125912007-125919468 (-)	PV	-	Prostate	Pro	MA	[15]
13q21.1	13:55627766-55635877 (+)	PV	-	Testis	-	MA	[15]
13q21.33	13:69795752-69799468 (+)	PV	LINC00383 (+) (Ex)	Testis	Pro°	MA, qRT-PCR	[83]
16p12.3	16:18124951-18125494 (-)	LTR	-	Testis	-	MA	[15]
17q22	17:53088886-53095859 (-)	PG	*STXBP4 (+)*	Testis	-	MA	[15]
21q21.1	21:20125060-20132866 (-)	PV	MIR548XHG (−) (Ex)	Testis	-	MA	[15]
21q21.3	21:28226756-28234297 (+)	PV	-	Testis	Pro°	MA, qRT-PCR	[15, 83]
Xq21.1	X:82517449-82517774 (-)	LTR	L1 PA11 (+), L1 MA2 (+)	Testis	-	MA	[15]
Xq23	X:113140352-113141135 (-)	LTR	L1 (-), XACT (-)	Testis	Pro°	MA, qRT-PCR	[83]
Xq24	X:120490096-120490859 (+)	LTR	-	Testis	PA	MA	[15]

[a] Chromosomal positions are referred to genome assembly GRCh37/hg19. Syncytin locus is highlighted in bold.

[b] PV= provirus, PG= processed pseudogene, LTR= solitary LTR.

[c] Colocalization of HERV-W element within a human gene: italic names correspond to coding elements, (Ex) indicates that the HERV-W sequence is colocalized with a gene exon.

[d] Tissues for which the HERV-W expression was reported also in physiological conditions are marked with a *.
[e] Reported activity of the sequence LTR: Pro= promoter, PA= PolyA signal. °Indicates a hypomethylated status with respect to normal samples.
[f] MA= microarray.

In particular, the majority of studies reported HERV-W sequences specific expression in tumoral testis, in line with the previously reported Syncytin-1 physiological expression in healthy testis [25], as well as in a number of other cancers affecting the genitourinary trait. Pérot and coworkers compared paired normal and tumoral tissues through a dedicated microarray, reporting a number of specific HERV-W loci differentially expressed in testis (16), prostate (2) and ovary (1) cancer samples [15]. Similarly, Gimenez and coworkers identified 6 specific HERV-W loci, including Syncytin-1 one, whose expression was upregulated in testicular cancer [83]. The comparison of these loci between normal and tumoral tissues revealed, in the latter, the general hypomethylation of U3 promoters, suggested to be a prerequisite for the transcriptional activation of at least five out of the six HERV-W loci [83]. In line with these findings, the analyzed HERV sequences global percentage of methylated CpGs was found to be greatly reduced (up to 30%) in tumoral tissues, even if some sequences were completely unmethylated in tumoral environments but not in the normal counterparts [83]. When considering bladder urothelial cell carcinomas, Syncytin-1 env was significantly hyperexpressed in >75% of the analyzed tumor tissues (n=82) compared to only the 6% of the matched adjacent tissues, increasing the proliferation and viability of immortalized human uroepithelial cells [85]. Interestingly, single nucleotide substitutions at positions 142 and 277 of Syncytin-1 3′LTR, were found in ~88% of tumoral tissues but only in a small proportion of healthy controls (~5%). Moreover, the 142T>C mutation was apparently driving the selective binding of c-Myb transcription factor to ERVWE1 LTR, being possibly associated to the Syncytin-1 promoter activity enhancement [85]. In addition to the above mentioned genitourinary tumors, HERV-W specific loci expression has been assessed also in mycosis fungoides, the most common type of Cutaneous T-Cell Lymphoma (CTCL) [84]. Two HERV-W loci in chromosomes 6 (6q21) and 7 (7q21.2), frequently harboring also abnormalities and rearrangements in CTCL, were predominantly and significantly upregulated in mycosis fungoides lesions as compared to the same patient intact skin [84]. In general, it is interesting to note that, beside the 9 HERV-W proviral sequences, also some L1-generated processed pseudogenes (6) and even solitary LTRs (8) were specifically upregulated in cancer tissues. This suggests that also defective

HERV-W elements, especially in the presence of an altered epigenetic control, can be actively transcribed and differentially expressed in cancerous tissues, possibly contributing to the disease progression.

Overall, until now, no human cancer was unequivocally related to HERV-W yet, as well as to any other HERV group, because of the lack of a definitive evidence that specific HERV sequences can induce tumors through the so far proposed mechanisms. In fact, although HERV expression in tumor tissues have probably some role in the disease clinical outcome, the presence of HERV RNA and proteins has not been demonstrated yet to be sufficient to support alone a causative role in tumorigenesis. The current results, instead, likely suggest that HERV-W group has a variable expression profile in both normal and cancerous tissues [86], even though the findings in both contexts are generally difficult to compare due to the use of a number of different experimental approaches [52]. In particular, as reported for the physiological context, also in tumoral tissues the lack of connection between the observed transcription profiles and the specific originating loci make it difficult to effectively assess any biological significance of HERV-W expressed elements. Also in this case, potential methodological biases could be associated to the use of Syncytin-1 provirus/MSRV cDNA sequences clones as a query and for the design of primers, due to the possible presence of HERV-W expressed loci with divergent or defective nucleotide sequence. Moreover, the lack of exhaustive information about HERV-W sequences basal expression in healthy tissues prevents the reliable evaluation of their effective altered expression in the correspondent tumoral contexts, the latter complicated also by an altered epigenetic regulation. In the light of this, even in the clear presence of a differential HERV-W expression between tumoral and healthy tissues, further studied are currently needed to establish which of such changes are actively involved in tumorigenesis and which other just constitute an epiphenomenon due to the altered tumoral environment [87]. Hence, important requirements for the proper assessment of HERV-W expression significance in tumors are the inclusion of both malignant and non-malignant tissues paired samples [84] and the evaluation of the methylation levels of LTRs as well as the genomic context of insertion of the investigated HERV-W sequences.

The current lack of a proven causative role for specific HERV-W loci in tumors etiology unfortunately represents a major obstacle for the exploration of HERV-based innovative anti-cancer approaches, possibly based on retroviral inhibitors, demethylating agents or RNA interference. Furthermore, also in the absence of a validated role in tumor etiology, the identification of HERV sequences selectively expressed in cancer tissues could provide a valuable target for passive and active anti-cancer immune therapies as well as specific biomarkers for the disease onset and progression.

1.4.3 Autoimmune diseases

Autoimmune diseases comprise a heterogeneous group of complex multifactorial disorders, sharing the incorrect recognition of healthy tissues and cells and/or the loss of immune tolerance to self Antigens (Ags) by the immune system. Clinically, such loss of tolerance leads to the development of B cells Antibodies (Abs) and/or cytotoxic T cells response to body components, which can end up in chronic inflammation and tissue destruction. A HERV role in the pathogenesis of autoimmune disorders was primarily suggested by the presence of retroviral Ags at the site of the disease and/or of specific Abs in the patient's sera [88, 89], and by an increased HERVs expression in patients with autoimmune disorders as compared to healthy individuals [90]. Theoretically, given that HERVs are stable components of the human genome, immune tolerance to such elements should have been established during development. Despite this, HERVs still show the evident ability to induce, or at least influence, both innate and adaptive immunity [88–93] (Figure 4). Currently, the most accepted theory is that HERV expression could evoke autoimmunity by molecular mimicry between common auto-Ags and exogenous retroviral proteins [90, 94–97]. HERV RNAs and proteins may, in fact, be recognized as PAMPs (Pathogen Associated Molecular Patterns) by innate immunity pathogen recognition receptors (PRRs) (recently reviewed in [93]), determining inflammation and the onset of auto-Abs produced by activated B lymphocytes at the site of disease. Moreover, HERV proteins may act as superAgs, triggering the non-specific polyclonal activation of auto-reactive T lymphocytes and inducing massive cytokine release. Beside the direct effect of retroviral immunogenic products, HERV proteins may deregulate the host immune response in alternative ways, e.g. by *trans*-activating/suppressing genes involved in immune modulation. In addition, even in the absence of any expressed product, the HERV mere presence could contribute to autoimmunity through insertional mutagenesis events and/or *cis*-regulation of nearby DNA, causing the transcriptional activation/inhibition/alternative splicing of immune regulatory genes.

Very recently, it has been also shown how (H)ERVs insertions dispersed a number of interferon (IFN) inducible enhancers in mammalian genomes, contributing in the evolution of a transcriptional network underlying the IFN pathway [98].

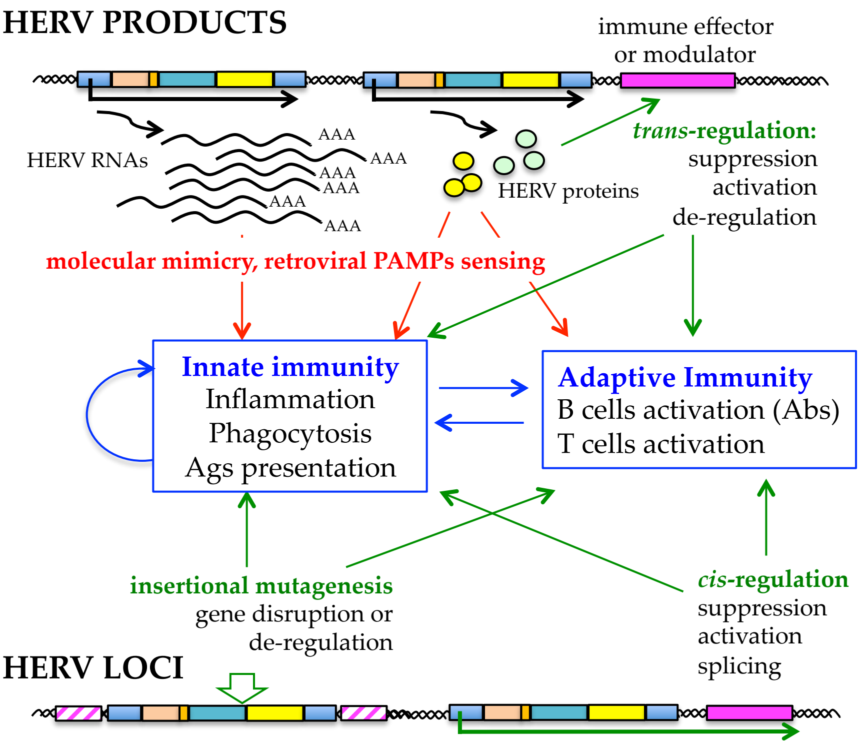


***Figure 4**. Potential mechanisms of HERV contribution to autoimmunity.*
HERVs can trigger autoimmunity by the direct sensing of their expressed products (red) as well as by mediating a immune effectors and modulators deregulation (green). In both cases, the eventual hypomethylated status at the site of autoimmunity could upregulate HERVs that are normally silenced in healthy tissues. HERV expressed RNAs and proteins (upper part) can act as PAMPs prompting the innate immunity effectors, and subsequently evoking an adaptive response. HERV proteins can act as superAgs activating a polyclonal expansion of autoreactive T cells, or can deregulate immunity effectors and modulators genes. These mechanisms can be also based on a molecular mimicry of HERVs products with respect to exogenous elements. HERV integrated sequences (lower part) can affect immunity also in the absence of expressed products, if their insertion disrupts or deregulates genes involved in immune response and its control. Also regulatory effects exerted by the sole HERV LTRs can alter the expression of such genes.

Importantly, as described for cancer, autoimmunity is also supposed to be influenced by the presence of an abnormal hypomethylation, which can eventually release the expression of HERVs normally silenced in healthy conditions [99]. Such phenomenon, on the one side, could contribute to the disease by providing nucleic acids or proteins acting as PAMPs, on the other side, it could be confounding, being possibly a consequence instead of an active contributor in the disease. In both cases, however, the loss of epigenetic control and the subsequent HERV hyperexpression can provide a high abundance of HERV-W transcripts suitable for the L1 *de novo* mobilization, as further possible contributors to the disease progression. For these reasons, also in autoimmunity, the proper identification and characterization of the specific HERV loci expressed in a precise pathological context is essential to assess their effective involvement in disease onset and persistence.

Focusing on the HERV-W group, the major field of investigation in autoimmunity is certainly MS, even if other autoimmune or immune-related disorders have been investigated by a few studies too, and are also reported.

1.4.3.1 Multiple Sclerosis

MS is an autoimmune disorder with poorly understood etiology, characterized by the progressive demyelination of the Central Nervous System (CNS). Both innate and adaptive immunity concur in MS immunopathogenesis, although adaptive immunity may predominate in disease onset, with selective T and B cell activation accompanying clinical relapses [100]. The precise causes of axons demyelination and damage remain unclear, although inflammatory molecules such as cytokines, chemokines, prostaglandins, reactive oxygen species and matrix metalloproteinases have been proven to contribute to MS-associated injuries [100]. Moreover, different infectious agents have been investigated for a possible link to MS [100–106].

As previously mentioned, the HERV-W group was initially related to MS due to its nucleotide identity to the so-called MSRV [107, 108], a putative retroviral elements detected in some MS patients samples either as virus-like particles or by the presence of RT activity [22, 37, 38, 109, 110], and proposed to be an exogenous competent member of the HERV-W group [22, 111–115]. The origin of MSRV is however still highly debated [116–118] and recent findings suggest that the published MSRV sequences could arise from the *in vitro* recombination of many HERV-W transcripts [21, 119]. In the last 20 years, a great amount of studies investigated the HERV-W/MSRV involvement in MS, mainly by i) the detection of HERV-W/MSRV nucleic acids in MS samples; ii) the presence of HERV-W/MSRV Ags in MS lesions; iii) the onset of an immune response against these elements, and iv) the use of some animal models of MS.

i) Regarding the presence of HERV-W/MSRV nucleic acids, a few studies investigated the differential amounts of integrated DNA sequences copy-number in MS samples, while most of them focused on the detection of expressed HERV-W/MSRV RNA transcripts. HERV-W/MSRV DNA copy-number was interestingly increased within MS patients PBMC as compared to controls, also correlating to disease severity [115, 120]. Considering that the group active proliferation ended several million of years ago, before the evolutionarily speciation of humans [121], it is unlikely that such variation could depend on the presence of unfixed proviral integration in a portion of the modern population, as shown for younger HERV groups. It is indeed more probable that, as described above, the additional HERV-W copies found in MS patients could be processed pseudogenes derived from novel L1-mediated retrotransposition events of HERV-W transcripts, specifically associated to the demethylated autoimmune environment. A positive relation between HERV-W/MSRV DNA copy number and female gender has been also hypothesized, possibly partially explaining the higher incidence of MS in women. In particular, the high number of HERV-W copies on X chromosome (12, of which 1 provirus and, remarkably, 10 L1-generated processed pseudogenes) could possibly play a role in MS sex-based variants, similarly to X chromosomes abnormalities [115]. Finally, in support to an endogenous origin of MSRV, MSRV *pol* sequences have been detected by fluorescence in situ hybridization (FISH) in the peripheral blood cells DNA from all patients with active MS, but they were found also in all healthy controls tested [122, 123]. Concerning the RNA expression level, HERV-W/MSRV *pol* sequences have been detected by RT-PCR approaches in MS patients brain [124]; leptomeningeal, choroid plexus and B cells [22]; peripheral blood lymphocytes [125]; cerebrospinal fluid (CSF) [22, 125, 126]; serum [126, 127] and plasma [125]. Overall, HERV-W/MSRV *pol* amplicons were found in a variable proportion of the analyzed MS samples (~50 to 100%) as well as in 0-50% healthy controls and 0-65% non-MS pathological samples, suggesting that HERV-W expression is not exclusive for MS but may be connected to the generic pathological environment and have a role in a particular subset of susceptible individuals. Unfortunately, the expressed RNA sequences were no attributable to specific HERV-W loci [102]. Also MRSV/HERV-W *env* RNA expression was reported to be upregulated in MS patients brain [128, 129] and PBMC [120]. Finally, a significantly higher accumulation of both HERV-W/MSRV *pol* and *env* RNAs was reported in MS brains [112] and CSF [130] with respect to healthy and pathological controls, even if all samples tested contained the HERV-W/MSRV transcripts regardless of the health/disease status.

ii) Similarly to HERV-W/MSRV nucleic acids, also the presence of HERV-W/MSRV proteins has been reported in both normal and MS brain tissues, questioning their effective role in MS pathogenesis. However, the presence of expressed proteins in diseased sites seems more likely to concur in the MS immune-pathogenicity and clinical manifestations. In fact, Syncytin-1 protein was present in MS patients brain and in specific cell types involved in lesions neuroinflammation, being indeed lowly expressed [128] or absent [120, 131] in controls. HERV-W Env epitopes were detected in higher quantities also on the surface of B cells and monocytes coming from patients with active MS with respect to stable MS patients and healthy controls, showing an increased seroreactivity [132]. Furthermore, Syncytin-1 *in vitro* expression mediated the production of proinflammatory molecules, potentially involved in astrocytes and oligodendrocyte damage [100], and an accumulation of HERV-W Gag Ags was shown in MS demyelinated brain lesions [133]. Finally, HERV-W/MSRV Env protein abundance in MS brain lesions was recently associated with areas of active demyelination, as predominantly expressed by macrophages and microglia, while moderate expression was observed in reactive astrocytes within active and chronic active lesions areas [134].

iii) In relation to HERV-W/MSRV Env, various and growing evidences suggest that such proteins may act as superAgs, triggering an abnormal innate immune response in MS, independently of a specific immune recognition pathway and resulting in the overproduction of cytokines, which are known to effectively play a major role in MS inflammatory demyelinating process. In details, MSRV Env induced, in both healthy donors and MS patients, the *in vitro* polyclonal activation of Vβ16 T-lymphocytes [135] and the subsequent increase of multiple pro-inflammatory cytokines [135, 136]. These HERV-W/MSRV Env pro-inflammatory properties have been attributed to the ability in triggering the Toll-Like Receptor 4 (TLR4) activation [135, 137, 138], leading to overexpression of the same proinflammatory cytokines known to play a role in MS inflammatory process, such as interleukins 1 and 6 (IL-1, IL-6) and Tumor Necrosis Factor α (TNF-α), and inducing a Th-1 lymphocyte polarization [136, 137, 139]. Moreover, HERV-W Env interaction with TLR4 and the subsequent upregulation of proinflammatory factors, in particular of the inducible nitric oxide synthase (iNOS), led to nitrotyrosine groups formation, which directly affected myelin protein expression and remyelination by blocking the oligodendrocyte precursor cell differentiation [140]. In line with this, the HERV-W Env neutralization by a specific monoclonal Ab (GNbAC1) reduced such stress reactions, rescuing myelin expression [141]. MSRV Env was also confirmed to be a potent agonist of human TLR4 *in vitro* and *in vivo* [142].

In addition to the Env superAg stimulation, some authors assessed the development of a specific immune response against HERV-W/MSRV in MS patients. Ruprecht et al. reported the presence of Syncytin-1 Abs in only 1/50 MS patients and in none of the 59 controls, whereas MSRV Gag or Env Abs were not detectable at all [118]. A similar result was obtained for the cytotoxic T-lymphocyte response [118]. In a follow up of MS patients Abs against HERV Env proteins, authors showed a marked (but not significant) decrease in anti-HERV-W Env reactivity as a consequence of Interferon (IFN) β therapy [143], as previously suggested also for circulating Env RNA [144]. A study assessing the humoral response against selected HERV-W Env peptides showed that two were strongly recognized by MS patients IgG as compared to controls, observing also a decrease of this specific response after 6 months of IFN-β therapy [145].

iv) Worth of note, the potential link between HERV-W/MSRV immunopathogenic properties and MS has been investigated also *in vivo* through some mice models, which generally supported an active involvement in the disease development. The intraperitoneal injection of MSRV virions in severe combined immunodeficiency (SCID) mice transplanted with human lymphocytes led to the onset of acute neurological symptoms, causing the animals death by massive brain hemorrhage [146]. Autoptic analysis confirmed the presence of circulating MSRV RNA and the overexpression of proinflammatory cytokines in spleen [146]. In another study, Syncytin-1 similarly induced neuroinflammation, neurobehavioral abnormalities and oligodendrocyte and myelin injury, principally evoked by redox reactant–mediated cellular brain damage [128, 147]. Always in mice, MSRV-Env was able to activate innate TLR4 and CD14 mediated proinflammatory cytokine release and, when associated to the myelin oligodendrocyte glycoprotein (MOG) Ag, to induce a specific T cells IFN-C production. Such combined innate and acquired responses promoted the development of experimental allergic encephalomyelitis, proposed as suitable MS model [148].

In addition to the high number of studies assessing the whole group general expression in MS, as for the other fields of HERV investigation, a limited number of studies were dedicated to the individual investigation of single transcribed HERV-W loci. The HERV-W processed pseudogene in locus Xq22.3 (ERVWE2) was among the most investigated, due to the presence of an almost complete *env* ORF, interrupted only by a premature stop at codon 39. Noteworthy, this L1-retrotransposed HERV-W element is transcribed in human PBMC [149–151], producing an N-terminally truncated Env protein (N-Trenv) *ex vivo* [152]. In addition, the evidence that a monoclonal Ab previously used to detect HERV-W Env in MS lesions (13H5A5) [133], was able to bind N-Trenv, but not Syncytin-1, allowed to speculate that this and other expressed defective proteins may exert some effects *in vivo* [152]. Also in line with the reported higher MS incidence in women, ERVWE2 locus in chromosome X has been proposed to be the hypothetical genomic origin of MSRV Env protein [115, 153] and investigated for its potential role in MS. However, the analysis of ERVWE2 DNA sequences in PBMC of MS patients and healthy individuals revealed that all of them harbor the above mentioned stop codon at site 39, assessing the absence of genetic polymorphisms which could possibly allow the production of a full-length protein *in vivo* [154]. Authors identified also 5 ERVWE2 DNA regions similar to the MOG Ig-like domain that, together with other co-factors, could trigger the immune cross reaction against myelin in MS [154]. García-Montojo et al. genotyped ERVWE2 insertion in a wide group of individuals, reporting the significant association of rs6622139 and rs1290413 polymorphisms to female MS susceptibility, being more frequent in controls than MS affected women [155]. A similar analysis was performed for an HERV-W insertion in chromosome 20, but the two identified polymorphisms resulted not significantly linked to different MS susceptibility based on case–control studies [156]. Some other works were aimed to assess the comprehensive HERV-W loci expression in MS. Laufer at al. tried to clarify the origin of the HERV-W/MSRV *env* sequences detected in MS samples by the evaluation of the HERV-W single loci expression. Interestingly, also the analysis of expressed HERV sequences was shown to be often complicated by *in vitro* recombination between HERV transcripts, likely due to RT template switches and/or PCR-mediated recombination [13, 151]. In particular, authors proposed that some previously published MSRV *env* sequences, as well as a high number of HERV-W env cDNA clones, were actually arisen from different HERV-W *env* transcripts recombination, detecting up to 4 recombination events involving up to 5 HERV-W loci for the same sequence. Moreover, it was shown how the commonly used primers could match to multiple HERV-W loci, underlining the importance of precisely assessing the detected transcripts genomic origin when studying HERV RNA expression [87, 151]. Of note, during the study, similar individual HERV-W *env* loci transcriptional levels were found in PBMC from MS patients and healthy controls [151]. Another comprehensive analysis of HERV-W loci brain transcription was performed by high-throughput sequencing of *env*-specific RT-PCR products, identifying >100 HERV-W loci transcribed at very similar levels in MS patients and healthy individuals [119]. Interestingly, while the deregulated expression of HERV-W *env* in MS brain lesions was consequently refuted, authors reported an inter-individual variability in HERV-W transcript levels, and a residual promoter activity for many HERV-W LTRs, even if incomplete [119]. Remarkably, a third study analyzing age- and disease-dependent

HERV-W *env* RNA diversity showed that HERV-W *env* transcripts were originated by multiple loci in primary human neurons; while astrocytes and microglia showed lower chromosomal diversity [157]. Similarly, while multiple loci encoding HERV-W *env* RNA sequences were detected in both fetal and adult healthy brain, transcripts cloned from neurologic patients mostly mapped to Syncytin-1 locus (7q21.2), and their abundance was highly correlated with pro-inflammatory gene expression in diseased brains [157]. This could indicate a wide and complex scenario, poorly clarified by the mere general HERV-W upregulation.

Taken together, the available reports about HERV-W/MSRV expression in MS patients do not definitively confirm a specific association of these retroviral elements to MS etiology yet, but strongly suggest a possible role of the group expression, especially at the protein level, in the disease immunopathogenesis. HERV-W/MSRV variable expression, as found in both MS patients and healthy individuals, could likely constitute a normal physiological phenomenon, possibly of higher prevalence in MS (and other diseases) due to an altered epigenetic and immunological environment [99, 158]. Moreover, the analysis of HERV-W/MSRV expression has been assessed mainly by RT-PCR, and the use of this technique for the detection and quantification of multicopy elements with an overall high homology has raised multiple doubts about its efficacy and specificity [100, 159, 160]. Also at the protein level, HERV-W/MSRV peptides showed variations according to ethnicity, opening the possibility of a role as co-contributors or predisposing factors, asking for additional studies about the HERV-W/MSRV brain proteomic profile in different ethnic populations [158]. Considering also the non-specific superAg activity of HERV-W/MSRV Env, showing neuropathogenic effects coincident with the major hallmarks of MS inflammation [134], the HERV-W group could possibly concur to a complex inflammatory interplay with other not fully understood factors, including genetic predisposition and exogenous infections [102, 161]. Noteworthy, a therapeutic treatment targeting HERV-W/MSRV has been proposed as a possible innovative approach for MS. GNbAC1 monoclonal Ab, developed to selectively recognize MSRV Env, after showing neutralizing effects *in vitro* and in MS mouse models is currently under phase II clinical development [162, 163].

1.4.3.2 Other autoimmune diseases

Beside MS, a few studies investigated the HERV-W group expression in other disorders with poorly understood etiology, in which autoimmunity mechanisms play a major pathogenic role. The main findings are related to Rheumatoid Arthritis (RA), Osteoarthritis (OA), Chronic Inflammatory Demyelinating Polyradiculoneuropathy (CIDP), Psoriasis and Lichen Planus (LP).

RA is characterized by the progressive destruction of the articular components, leading to severe disability. A common sign of autoimmune response is the presence of a cellular infiltrate of neutrophils, lymphocytes and macrophages in the synovial tissue, accompanied by the increased production of metalloproteinases contributing to the extracellular matrix erosion [164]. Based on preliminary results reporting HERV-W/MSRV RNA in the 50% of RA patients plasma samples, Gaudin et al. investigated the presence of particle-associated HERV-W/MSRV RNA, as indicative of virion production [164]. Results showed that neither the patients nor the controls had HERV-W/MSRV RNA in plasma, while such RNA was detected in synovial fluid samples of 2/9 RA patients and 1 control, suggesting however its lack of specificity with respect to RA etiology [164].

OA is another common form of arthritis characterized by the progressive destruction of articular cartilage, in which many factors, including viral infections, seem to play a role [165]. Bendiksen et al. analyzed cartilage and chondrocytes from advanced-OA and early/non-OA: while all samples were negative for a number of exogenous infections (including parvovirus B19, Herpesviruses and hepatitis C virus), an *env* gene of the HERV-W group was commonly expressed in advanced OA patients cartilage (88% of patients) as well as in a proportion of controls (0-38%) [165]. Authors reported also abundant expression of Env proteins in OA-derived chondrocytes, and the occurrence of viral budding and virus-like particles that were however neither isolated nor characterized [165].

Another pathology tentatively linked to HERV-W/MSRV is CIDP, a rare immune disease of the peripheral nervous system characterized by inflammatory and demyelinating lesions in nerve roots [166]. Driven by the presence of MSRV-Env in a little number of CIDP patients (5/8) included as pathological non-MS controls in a previous study [120], Faucard et al. confirmed an upregulation of MSRV *env* and/or *pol* mRNAs in ~50-65% of CIDP patients PBMC as compared to controls, which presented lower percentages of positivity [166]. Authors reported also the presence of MSRV Env protein in 5/7 CIDP patients nerve lesions (but not in the 2 control tissues), showing dominant expression in Schwan cells [166]. Moreover, Schwan cell cultures exposed to MSRV-Env displayed a potent induction of IL-6 and CXCL10 chemokine, significantly inhibited by GNbAC1 MSRV-Env mAb [166].

Finally, also some skin diseases were suggested to involve HERV-W/MSRV expression in their unclear etiology. Psoriasis is a chronic disease characterized by epidermal proliferation and abnormal keratinocytes differentiation, and shows also systemic immunological disorders closely related to autoimmunity [167]. Considering that HERVs expression has been reported in human skin, being either activated or repressed by UV irradiation [168, 169], Molès et al. assessed HERV expression in psoriatic lesions, showing the presence of various *pol* sequences, comprising HERV-W *pol* amplicons, in both psoriatic and control skin [167]. Another pathology taken into account was LP, a skin chronic inflammatory disease characterized by lichenoid papules and possibly involving also viral and bacterial agents in its unclear etiology [170]. De Sousa Nogueira et al. observed a downregulation of some HERV groups, including HERV-W *env*, in skin biopsies of LP patients, with a concomitant activation of antiviral restriction genes (APOBEC3G, MxA and IFN-inducible genes) hence possibly involved in the immune control of HERVs transcription [170].

1.4.4 Neurological and neuropsychiatric disorders

Also in the light of the findings in the context of MS, the HERV-W neuropathogenic effects have been investigated in a number of neurological and neuropsychiatric diseases with poorly understood etiology, i.e. Motor Neuron Disease (MND), sporadic Creutzfeldt–Jakob Disease (CJD), Autistic Spectrum Disorder (ASD), Attention Deficit Hyperactivity Disorder (ADHD) and Schizophrenia.

MND is a heterogeneous group of neurologic disorders characterized by progressive degeneration of motor neurons. Elevated levels of HERV-W *env* transcripts were observed in biopsies from MND patients limbs as compared to control tissues from the same individual and from healthy donors [171]. Authors detected also a parallel upregulation of the SOD1 (oxidative stress-responsive) gene, a marker for oxidative stress, suggesting that its activation could constitute a response of the degenerating muscle fibers to the primary loss of motor neurons instead of a consequence of HERV-W Env neurotoxic effects, being accompanied by a subsequent infiltration of phagocytic cells expressing HERV-W in the atrophic muscle [171].

Sporadic CJD is a rare form of prion disease causing fatal neurodegeneration and having as key event the conformational change of cellular prion protein to an abnormal protease-resistant isoform. Joang et al. examined the expression of 10 HERV groups in sporadic CJD patients CSFs, detecting transcripts for all the groups taken into account and reporting the highest incidence for HERV-W *pol* (82,5% positivity), with a significant increase (as for HERV-L, -FRD and -9) with respect to controls [172]. Based on subsequent subcloning analysis, all the observed transcripts showed unidentical nucleotide sequence, and none had specificity for sporadic CJD [172].

ASD and ADHD are two neurodevelopmental diseases caused by complex interactions among not fully clarified genetic and environmental factors. ASD patients PBMC showed higher positivity for HERV-H and HERV-W mRNAs as compared to controls [173]. Subsequent quantification showed that HERV-H and HERV-W were more and less abundantly expressed in ADS patients, respectively [173]. Similarly, the HERV transcripts amount in ADHD patients PBMC was significantly higher for HERV-H, while no differences were found in HERV-W expression [174].

Among neuropsychiatric disorders, the field of greatest interest for HERV-W potential involvement is schizophrenia. The first findings about a possible HERV-W contribution were provided by Deb-Rinker et al. in monozygotic twins discordant for schizophrenia, presenting one sequence (schizophrenia associated retrovirus, SZRV-1, AF135487) similar to both a MSRV (AF009668) and a HERV-9 (S77575) sequences expressed in placenta [175]. Karlsson et al. detected then HERV-W/MSRV *pol* sequences in the cell-free CSF from ~29% acute onset schizophrenia patients and 5% individuals in later stages of the disease, but not in patients with non-inflammatory neurological diseases and healthy controls [176]. Similarly, HERV-W/MSRV expression was up-regulated in the brain frontal cortex regions of schizophrenia patients as compared with corresponding control tissues from healthy individuals [176]. In subsequent studies, the same authors reported the presence of HERV-W RNA in the plasma of a subgroup (9/54) of recent-onset schizophrenia patients, 5 of which harbored HERV-W/MSRV sequences in CSF [177], and observed also elevated total levels of HERV-W *gag* (but not *env*) transcripts in PBMC of patients with schizophrenia-related psychosis, reporting an upregulation of HERV-W sequences in locus 11q13.5 [178]. HERV-W *env* plasmatic mRNA was found in 36% of recent-onset schizophrenia patients and in none of the 106 controls, and also RT activity was significantly increased in patients sera [179].

At the protein level, HERV-W Env hyperexpression in U251 human glioma cells triggered the production of the dopamine receptor D3 and the brain-derived neurotrophic factor (BDNF), both associated to schizophrenia, and increased the phosphorylation of CREB protein, necessary for BDNF expression [179], as confirmed also in human neuroblastoma cells [82]. Moreover, recent findings suggested that also the phosphorylation of Glycogen Synthase Kinase 3β might be involved in HERV-W Env-mediated BDNF induction [180]. A study detecting HERV-W Ags in living patients reported positive serum antigenemia for Gag and Env in ~50% of schizophrenic patients and in 3-4% blood donors [181]. Of note, a full-length HERV-W LTR was found in the regulatory region of GABBR1 (GABA receptor B1) gene, which is downregulated in schizophrenic patients [182]. However, the roles of this LTR and GABBR1 in schizophrenia remain to be clarified.

In contrast to these studies showing an increased HERV-W expression, at least in a subgroup of schizophrenic patients [177–179], a number of investigations presented instead an opposite scenario, reporting no specific correlation between the group transcription and the development of neurological diseases [183–185]. The comprehensive microarray-based analysis of 20 HERV groups transcriptional activity in 215 brain samples from schizophrenia or bipolar disorders (BD) patients and matched controls did not show relevant links between HERV brain transcription and schizophrenia, being more likely influenced by the individual genetic background and the presence of immune cells infiltrates and/or medical treatments [183]. Interestingly, different brain areas of each individual showed a common pattern of HERV expression, where HERV-W *env* gene was found to be transcriptionally active but did not show significant differences between healthy controls and schizophrenic patients [183]. Weis et al. observed that HERV-W Gag proteins are physiologically present in human brain anterior cingulate cortex and hippocampus, mostly associated to neurons and astrocytes, showing a significantly reduced expression in schizophrenia, major depression, and BD patients as compared to controls [184]. HERV-W *env* transcription was increased in schizophrenia and BD patients PBMC, but the corresponding DNA copy number was paradoxically significantly lower in patients than healthy controls. Moreover, differences in HERV-W *env* amplicons nucleotide sequences and their relative frequencies were observed comparing patients to controls and between Schizophrenia and BD patients and MS [185]. Authors hypothesized that during development (when HERV-W genes are hypomethylated), an environmental stimulus (such as exogenous infections) could prompt lineage-specific HERV-W genomic modifications and constitute variable patterns that would respond differently to subsequent environmental triggers, leading to diverse clinical manifestations [185].

In summary, the available information does not definitively support yet a direct causative role of HERV-W group in any neurological or neuropsychiatric disease. Of note, in fact, a proportion of HERV-W-negative patients is reported in the majority of the studies, while a significant upregulation of HERV-W expression was shown in a subset of cases, strongly suggesting the presence of other major factors contributing to a complex and poorly understood etiology. It is also worth to note that many of these pathologies could be concomitant with behavioral variables, such as drug and alcohol abuse [184], which can be confounding if able to influence HERV brain expression. In particular, a study investigating the effects of antipsychotic drugs on HERVs transcription in brain suggested that their intake may contribute to augmented and differential expression of distinct HERV groups in neuropsychiatric patients, reporting a significant increase in HERV-W transcription associated with valproic acid treatment in schizophrenic individuals [186]. Differently, no influence of alcohol or drug consumption on HERVs brain expression was detected [186].

1.4.5 Infectious diseases

HERVs have been also proposed to have a role during exogenous viral infection, and such role could be either beneficial or maleficent. On one hand, HERVs antisense transcripts have been hypothesized as plausible defense mechanism against exogenous retroviral infections, in which the complementary interaction between homologous RNA sequences could form dsRNA, a known PAMP detected by the innate immunity effectors PRRs [93, 187]. Another HERV-mediated antiviral effect could be a partial resistance to infection, evoked by receptor interference and blocking by HERV proteins [188, 189]. On the other hand, exogenous viruses and expressed HERVs may also generate cooperative effects, stimulating each other transcription or leading to complementation of defective elements. Clearly, some of these interactions require a certain degree of sequence and structural homology, and are most likely to happen between HERVs and exogenous Retroviruses.

1.4.5.1 Retroviral infections

Humans are currently threatened by two exogenous retroviruses: Human Immunodeficiency Virus (HIV, *Lentiviridae*) and Human T-cell Lymphotropic Virus (HTLV, *Deltaretroviridae*). HERV-W Env glycoprotein was shown to functionally complement an *env*-defective HIV-1 strain, generating HERV-W-pseudotyped particles infectious for CD4-negative cells, as possible and interesting mechanism of HIV-1 tropism expansion [190]. HERV-W elements were upregulated in three persistently HIV-1 infected cell lines, but not in *de novo* infected cells [191]. Interestingly, the treatment of HIV-1 latency T cells with histone-deacetylase inhibitors, known to reactivate latent HIV-1 in clinical trials, caused no substantial increases of HERV-W *env* gene transcription [192]. Looking for a HERV-W contribution to AIDS neuropathy, a significant HERV-W RNA hyperexpression was detected in brain from patients suffering of AIDS dementia, but the observed variation was a consequence of increased immune activity, linked to monocyte differentiation and macrophage activation [124]. In this regards, HIV Tat transactivator protein increased MSRV *env* mRNAs and HERV-W Env protein expression in astrocytes and differentiated macrophages, showing instead reduced expression in monocytes [193]. Similarly, HTLV-I Tax, homologous of Tat, is likewise also able to activate HERV-W LTRs by interacting with many transcription factors, including CREB [194]. T cell cross-reactivity between HERVs and HIV epitopes was tested *in vitro*, giving negative results [195].

1.4.5.2 Herpesviral infections

The possible interaction between Herpesviruses and HERV-W expression has been widely analyzed, especially in the context of MS and other autoimmune diseases [89–91].

HERV-W/MSRV expression was enhanced by Herpes Simplex Virus 1 (HSV-1) superinfection in MS patients cells [37, 101, 196]. More in details, Lafon et al. showed that HERV-W Env proteins expression in neuroblastoma cell lines can be reactivated by HSV-1, probably through its ICP0 and ICP4 proteins [101]. HERV-W Gag and Env proteins were also induced by HSV-1 in neuronal and brain endothelial cells *in vitro*, and the observed expression was similarly compatible with an ICP0-mediated activation [197]. Other evidences have been reported in HeLa cells, in which HSV-1 IE1 protein stimulated the LTR-directed transcription of HERV-W, probably through the modulation of Oct-1 cellular transcription factor [198]. Authors proposed that IE1 activation could possibly interests also HERV-W solitary LTRs, potentially promoting eventual nearby genes [198].

Beside HSV-1, also other Herpesviruses have been investigated in relation to HERV-W activation in MS. A small hypothetical ERVWE1 Env peptide (29 aa) harbors an epitope predicted to be presented by different HLA class I molecules, and possibly acting as a target for effector T-cells in MS. Interestingly, such epitope showed homologies in all the pathogens against which elevated Abs titers were found in MS patients, including HSV-1, HHV-6, VZV (Varicella Zoster Virus) and EBV (Epstein Barr Virus) as well as measles virus [199]. Hence, it was claimed that the effector T cell recognizing this putative epitope would most readily cooperate with regulatory T cells to support a protective immune response, leading either to a prompt resolution of the infection, or to tissue damage by autoimmune processes [104, 199]. Regarding EBV, the exposure to the virus or its major Env glycoprotein (gp350) triggered the HERV-W/MSRV expression in PBMC from MS patients and MSRV positive healthy controls as well as in cultured U87-MG astrocytes, with an activation pathway possibly involving NF-kB [200]. Infection of various cancer and non-cancer cell lines with CMV induced RT activity in all cells, determining also the upregulation of various HERVs, including HERV-W, in CMV-infected neural tumor stem cells also after UV irradiation [201]. Other evidences of a CMV helper role in HERV-W activation came from kidney transplant recipients with high CMV load, showing significantly higher HERV-W *pol* expression levels than the ones found in groups with no or moderate CMV load and as compared to healthy subjects [202].

1.4.5.3 Other exogenous infections

Despite Retroviral and Herpesviral infections are the most intensively studied for their effects on HERV-W expression, also Influenza, Spleen Necrosis Virus (SNV) and Porcine Endogenous Retrovirus (PERV) infections have been implicated in HERV-W modulation.

Nellåker et al. described specific expression patterns of HERV-W *gag* and *env* sequences, constitutively transcribed in different cell-lines also if harboring truncated/no LTRs, observing subsets of elements being transactivated by influenza active replication, while not by an indirect consequence of antiviral response mechanisms [149]. However, similar variations were observed as a consequence of serum deprivation, suggesting that also cellular stress could contribute to HERV-W modulation [149]. Subsequent analysis showed that Influenza infection induces spliced ERVWE1 transcripts able to encoding Syncytin-1 in extra-placental cells by GCM1 overexpression [203] and the downregulation of the repressive histone mark H3K9me3 in regions harboring HERV-W elements [204].

The HERV-W Env glycoprotein induced cellular resistance to SNV, whose infectivity was reduced by 1000- 10,000-fold in D-17 cells expressing HERV-W Env [205].

Finally, in the field of xenotransplantation, the expression level of HERV-W genes differed in PERV-infected HEK-293 cells in comparison to uninfected cultures [206].

1.4.6 Conclusions

In all the pathological contexts investigated for a HERV-W etiological role, the great majority of studies detected and quantified the entire group overall expression, often through not fully standardized methodologies on differently representative samples, producing a great amount of data, unfortunately frequently discordant and difficult to compare one to each other. On the one side, the currently available information strongly suggest that the HERV-W group is commonly transcribed in human cells, in both healthy and pathological conditions, showing a collective expression that greatly vary between tissues and according to the individual genetic background. On the other side, the systematic evaluation of HERV-W individual sequences transcription has never been performed yet, preventing until now a strongly reliable association to human diseases. Thus, the yet to be explored expression pattern of single HERV-W loci will be essential to provide more insights on the quantitative changes originated from specific HERV-W sequences instead of detecting the overall group expression, and thus to identify those single members possibly linked to human pathogenesis. Even if the available information strongly suggests a link between the HERV-W expression and many of the pathological context analyzed, especially regarding cancers and autoimmune diseases, this promising field deserves a deeper investigation to characterize the many aspects still poorly understood, and to explore all the possible mechanisms involving HERV-W presence and expression in both physiological and pathological conditions. The latter includes also the possibility that an altered epigenetic environment could prompt the *de novo* mobilization of HERV-W transcripts by the L1 elements still active in the human genome, providing additional HERV-W processed pseudogenes insertion with respect to the ones recently mapped in the human genome reference sequence [121].

The current poor knowledge of the individual HERV-W loci transcriptional status is in big part due to the previous absence of a comprehensive HERV-W genomic characterization, leading to design of all experimental procedures on a few HERV-W sequences, above all Syncytin-1 and MSRV clones. This prevented until now i) the univocal assignment of the reported HERV-W expressed sequences to the locus of origin; ii) the characterization of the single HERV-W sequences differential expression in the diverse physiological tissues, fundamental to assess the effective dysregulation in diseased environment; iii) the evaluation of the full-length HERV-W sequence transcription, coding capacity and regulatory elements; iv) the characterization of the single HERV-W sequences epigenetic status in both physiological and pathological contexts, and iv) the study of the HERV-W sequences genomic context of insertion and of the presence of eventual nearby host genes potentially influenced by HERV-W elements, even in the absence of a detectable expressed product.

For future research investigating HERV-W contribution to human physio-pathology, dedicated genome-wide studies as well as stringent primers and probes, able to distinguish the uniqueness of single HERV-W elements conforming to standard conditions, are needed to properly define the specific contribution of the different retroelements to the human transcriptome [100]. Importantly, the physiological single HERV-W loci expression must be evaluated *a priori*, in order to have a reliable "basal" level to compare with the same, diseased, tissue [90]. Such specific quantitative analyses must be then performed on a statistically significant population, possibly including paired samples of both healthy tissues and pathological lesions from the same individual. Moreover, these investigations need to take into account also the influence of the HERV-W loci genomic context of integration, and to analyze the single insertions molecular diversity within the human population, due to the possibility that different HERV-W allelic variants may exert specific effects on the pathogenesis phenotype, progression and therapeutic response, depending on the host genetic background [100]. In the case of HERV-W, as mentioned above, considering that ~80-100 copies of L1 are estimated to be still active in the human genome [35, 207, 208], their mediated retrotransposition of HERV-W sequences should also be considered, especially in those pathological environments associated to an altered epigenetic control, where the hypometilation of both L1 and HERV-W sequences have been reported [75]. Finally, considering that also structurally incomplete LTRs could be still able to drive the transcription of HERV-W proviruses, processed pseudogenes or nearby host genes, also the methylation levels of truncated and solitary LTRs should be evaluated. Indeed, beside the mere detection of HERV-W group overall expression, the identification of the specific encoding locus appears mandatory to establish any definitive associations between human diseases and specific retroelements, and also to properly understand the molecular nature of emergent forms arisen by recombination events involving different HERV-W loci [100], especially in those context where the epigenetics alteration could liberate HERVs expression. Similarly, also the molecular determinants responsible for specific HERV-W loci upregulation as well as their relation as a cause or a consequence of disease must be clarified in detail [102], finally providing a well-characterized mechanism of HERV-mediated pathogenesis.

Overall, based on the current knowledge, it is surely possible that specific HERV-W sequences may play a role in human pathogenesis, without necessary being the only etiological determinant of disease. More likely, especially in the field of autoimmunity, one or more HERV-W insertions (or even a specific allelic variant) and/or their expressed products could be involved in a complex inflammatory and immune interplay with other unknown or not fully understood co-factors. The latter may include individual predispositions, depending on the host genetic background, as well as extrinsic factors such as stress, environmental stimuli or exogenous infections. All these complex relationship must be considered, especially in the field of multifactorial disorders with poorly clarified etiology.

In conclusion, the identification and characterization of the precise HERV-W loci showing a differential transcription pattern and/or L1-mediated HERV-W *de novo* mobilization in a specific pathological context appears mandatory to definitively demonstrate a cause-effect connection to any disease etiology, and to subsequently identify single HERV-W sequences as novel therapeutic targets. The latter could be suitable to various, innovative approaches, from the employment of retroviral inhibitors to the administration of passive as well as active immunotherapy directed against specific HERV products, also in association to the treatment with DNA-demethylating agents. However, also in the absence of an etiological contribution, the identification of specific HERV-W sequences selectively expressed in a given pathological context could provide novel and reliable biomarker of diseases, or disease-associated Ags also suitable to direct immunotherapeutic approaches to the precise site of pathogenesis. All these innovative HERV-based therapeutic applications could surely constitute an innovative treatment for human diseases, and this still unexplored possibility strongly require studies aimed to definitively assess the HERV specific contribution to human physio-pathology.

1.5 AIM OF THE WORK

Although TEs have been considered as mere genomic parasites for a long time, the evidence that such a wide proportion of eukaryotic DNA is composed of mobile elements suggests that their presence could not be only detrimental to the fitness of the host [209].

In this context, HERV-W group provides coding elements to perform specialized tasks in the placenta and is one of the most remarkable examples of TEs exaptation for an essential physiological role. Moreover, the group generic expression has been tentatively involved into a wide number of disorders, but the absence of any proved mechanism of pathogenesis of a specific HERV-W sequence prevented until now the definitive connection to any human disease. Also considering that the human genome is estimated to harbor 80–100 L1s still active and competent for retrotransposition [208, 210, 211], the unique capacity of HERV-W expressed RNA to be mobilized by L1 elements could still represent an indirect source of ongoing reinsertions, which further contribute to intra- and inter-individual genetic variation and, on occasion, to sporadic genetic disorders [208, 210, 211].

Hence, the aim of this work is the comprehensive and updated characterization of the HERV-W group sequences integrated into the human genome, as well as in the non-human primate genomes, establishing an unique database useful to connect the observed expression profiles to the uniqueness of each HERV-W loci. The detailed analysis of HERV-W proviruses and processed pseudogenes in terms of sequences, structural features, phylogeny, time of integration and genomic context of insertion will allow to further clarify the role of these elements in human physiopathology. Moreover, the analysis of the group single members in the humans ancestors will tentatively define the dynamics of HERV-W group acquisition up to humans, possibly highlighting unreported insights about its evolution, occurred in parallel to primates speciation.

Chapter 2. Materials and Methods

2.1 HERV-W sequences identification and localization

2.1.1 HERV-W sequences in humans

The identification and collection of HERV-W sequences has been done in human GRCh37/hg19 assembly (released in February 2009) using a double approach that binds i) the hg19 assembly FASTA sequence analysis by the RetroTector program package (ReTe) [212] and ii) a traditional BLAT search [213] in the UCSC Genome Browser database [214] using the RepBase Update [215] assembled LTR17-HERV17-LTR17 consensus as a query. The elements found by both approaches have then been confirmed as HERV-W based on i) Repeat Masker analysis of the HERV-W sequence and its genomic flanking portions, ii) structural alignment and visual comparison with respect to the HERV-W group RepBase reference LTR17-HERV17-LTR17 and iii) phylogenetic trees of LTRs and *gag*, *pol* and *env* genes; in order to avoid misclassifications or incomplete sequences inclusion. The retrieved HERV-W elements were named according to the locus of genomic localization, and in the presence of multiple sequences in the same locus, the order within the band was expressed with a letter following the alphabetical order.

HERV-W solitary LTRs were retrieved by UCSC Genome Browser BLAT search using LTR17 as a query, and kindly provided by Professor Jens Mayer (Saarland University).

2.1.2 ERV-W sequences in non-human primates, Catarrhini parvorder

The identification and collection of ERV-W sequences was performed in the following *Catarrhini* primates genomes assemblies on Genome Browser database [216]:

- Chimpanzee (*Pan troglodytes*, assembly Feb. 2011 - CSAC 2.1.4/panTro4)
- Gorilla (*Gorilla gorilla gorilla*, assembly May 2011 - gorGor3.1/gorGor3)
- Orangutan (*Pongo pygmaeus abelii*, assembly July 2007 - WUGSC 2.0.2/ponAbe2)
- Gibbon (*Nomascus Leucogenys*, assembly Oct. 2012 - GGSC Nleu3.0/nomLeu3)
- Rhesus (*Macaca Mulatta*, assembly Oct. 2010 - BGI CR_1.0/rheMac3)

In particular, the identification and collection of ERV-W sequences orthologous to previously characterized HERV-W loci was done by comparative localization of the correspondent human genomic region in the analysed non-human primates genome sequence. The presence/absence of each ERV-W orthologous locus was assessed by including in the examination a minimum of 500 nt of the 5′ and 3′ flanking genomic sequence, that are shared among the different primates species and ensured thus to precisely localize the corresponding chromosomal position within the orthologous genome region.

Moreover, in order to confirm the previously localized ERV-W sequences as orthologous to human loci and to retrieve also eventual additional ERV-W sequences in non-human *Catarrhini* primates lacking a correspondent orthologous insertions in humans, the above mentioned non-human primates genomes sequence assemblies were analysed by BLAT searches [213] using an assembled sequence consisting of LTR17-HERV17-LTR17 as provided by RepBase Update [215] as a query.

Each ERV-W locus identified by BLAT in non-human primate genome sequences was mapped to the human genome to investigate presence of orthologous ERV-W elements by using comparative genomics data as provided by UCSC Genome Browser and considering a minimum of 500 nt of 5′ and 3′ flanking genomic sequence for each investigated locus. Absence of a HERV-W sequence in an orthologous genome region was concluded when no HERV-W sequences i) were found in that genome region by BLAT searches using HERV17 as a query, ii) were found at the correspondent orthologous position by flanking sequences comparative localization and iii) were found by BLAT search using the ERV-W nucleotide sequence from the respective orthologous primate genome region as a query.

2.1.3 ERV-W-like sequences in non-human primates: Platyrrhini parvorder

A number of ERV-W-like elements were collected by a UCSC Genome Browser BLAT search, using the RepBase HERV17 sequence as a query, in the following *Platyrrhini* (*Cebidae* family) primates genome sequence assemblies:

- Marmoset (*Callithrix jaccus*, assembly March 2009 - WUGSC 3.2/calJac3)
- Squirrel Monkey (*Saimiri boliviensis*, assembly Oct. 2011 - Broad/saiBol1)

Retrieved ERV-W-like sequences were pre-annotated by Genome Browser Repeatmasker/RepBase as ERV1-1_CJa-I and ERV1-1_CJa-LTR regarding internal portions and LTRs, respectively. ERV1-1 sequences were downloaded in FASTA format including 500 nucleotides of 5′ and 3′-flanking. Sequences harbouring relatively intact LTRs, based on pairwise dot-plot comparison, were selected for subsequent analysis.

Since no assembled genomes sequences are available for representative member of the other two *Platyrrhini* families, *Atelidae* and *Pitheciidae*, the presence of ERV-W-like elements in these primates was assessed by BLAST searches of unassembled genomic sequences data available from the NCBI Trace Archive database (https://trace.ncbi.nlm.nih.gov/Traces/sra/) for:

- Spider Monkey (Ateles geoffroyi, *Atelidae* family – Trace Archive),
- Red-bellied Titi (Callicebus moloch, *Pitheciidae* family – Trace Archive)

using LTR17-HERV17-LTR17 and a majority rule generated proviral consensus of ERV1-1 as queries.

2.1.4 ERV-W-like sequences in non-human primates: Tarsiiformes and Prosimians

ERV-W-like elements were searched by BLAT searches at the UCSC Genome Browser using LTR17-HERV17-LTR17 and a majority-rule proviral consensus of ERV1-1 as queries in the following *Tarsiiformes* and *Prosimians* genome sequence assemblies:

- Tarsier (*Tarsius syrichta, Tarsiiformes*, assembly Sep. 2013 – Tarsius_syrichta-2.0.1/tarSyr2)
- Bushbaby (*Otolemur garnettii*, *Lemuriformes*, assembly Mar. 2011 – Broad/otoGar3)
- Mouse Lemur (*Microcebus murinus*, *Lorisiformes*, assembly Jul. 2007 – Broad/micMur1)

2.2 Sequences pairwise and multiple alignments

Pairwise and multiple alignments of nucleotide and amino acid sequences were generated using Geneious bioinformatics software platform, version 8.1.4 [217] using MAFFT algorithms FFT-NS-i x1000 or G-INS-i [218] with default parameters. All alignments were visually inspected and, when necessary, manually corrected before subsequent analysis. Pairwise comparisons of sequences were done using the dot-plot analysis tool implemented in Geneious. The alignments graphical representations were also obtained from Geneious bioinformatics software.

2.3 HERV-W sequences structural characterization

The GRCh37/hg19 assembly retrieved HERV-W sequences were multiple aligned with respect to the RepBase Update assembled LTR17-HERV17-LTR17 reference, and each sequence nucleotide composition was characterized in detail with respect to the RepBase Update assembled LTR17-HERV17-LTR17. To this purpose, all insertion and deletions ≥ 1 nucleotide were annotated, and the presence of other repetitive elements as secondary insertion was also reported.

2.4 Phylogenetic analysis

2.4.1 Phylogenetic trees

All phylogenetic trees were built from manually optimized multiple alignments generated by Geneious (see above) using Mega Software, version 6 [219] applying either Neighbor Joining (NJ) and/or Maximum Likelihood (ML) statistical methods. In particular:

- ❖ For nucleotide alignments: NJ trees were built using p-distance or Kimura 2-parameter models and applying pairwise deletion treatment, phylogeny was tested by bootstrap method with 500 or 1000 replications. ML trees were built using Kimura 2-parameter model, phylogeny was tested by bootstrap method with 50, 100 or 1000 replications.
- ❖ Amino acids alignments: NJ trees were built using p-distance or Poisson correction models and applying pairwise deletion treatment, phylogeny was tested by bootstrap method with 500 or 1000 replications. ML trees were built using Poisson correction model, phylogeny was tested by bootstrap method with 50, 100 or 1000 replications.

2.4.2 Nucleotide and amino acids distances

The pairwise divergence between nucleotide or amino acids sequences was estimated on visually-inspected alignments, after the removal of hypermutating CpG dinucleotides, by using Mega Software, version 6 [219], and a p-distance model applying pairwise deletion option.

2.5 Time of integration estimation

The age of each HERV-W and ERV1-1 sequence was estimated through different approaches, all based on the calculation of the percentage of nucleotide divergence (D).

In particular, the D values were estimated on Mega software (version 6) by Kimura 2-parameter corrected pairwise distances excluding gaps and CpG dinucleotides, between: i) the 5′ and 3′LTRs of each provirus, ii) each HERV-W/ERV1-1 sequence (proviruses and processed pseudogenes) single LTRs and a generated consensus for each LTR subgroup, and iii) each HERV-W/ERV1-1 sequence (proviruses and processed pseudogenes) *gag* gene (ERV1-1) or a 150-300 nucleotides portion of *gag*, *pro*, *pol* RT, *pol* IN and *env* genes (HERV-W) and a generated consensus for each group.

The obtained D values have then been used, according to previous methodologies [220], to estimate each HERV-W/ERV1-1 sequence time of integration (T), based on the relation

$$T = \frac{D}{SR}$$

where SR is the estimated host genome random substitution rate percentage. In details:

- For HERV-W sequences T estimation, the assumed human genome SR was 0,13%/nucleotides/million years. Thus, the calculation was T=D/0,13.
- Regarding ERV1-1 sequences T estimation, a univocal SR for *Platyrrhini* coding regions is currently not available in literature, and no estimated SR for non-coding regions has been defined yet. Thus, we decided to perform the initial age estimation considering i) two of the proposed *Platyrrhini* genomic SR reported in literature (0,14% and 0,126%) [221, 222], ii) a SR recently used by our group to estimate HERV proviruses age (0,2%) [8] and iii) an averaged SR obtained from these values (0,16%).

For both HERV-W and ERV1-1 proviruses, the T values obtained from 5′ and 3′LTRs D calculation were further divided by a factor of 2, assuming that each LTR evolved independently into the genome (T=D/SR/2). The final age of each sequence was expressed as the average of the T values obtained with the different approaches, excluding those values with a standard deviation > 20%.

2.6 PBS type and Gammaretroviral features analysis

All HERV-W and ERV1-1 elements were aligned and analyzed for the presence of i) PBS nucleotide sequence, ii) Gag NC Zinc finger amino acid motif, iii) Pol IN C-terminal GPY/F amino acid motif and iv) any bias in the overall nucleotide composition along the sequence.

The degree of conservation of PBS nucleotide sequence as well as Gag NC Zinc finger and Pol IN C-terminal GPY/F amino acid motifs was represented through a logo built from WebLogo at http://weblogo.berkeley.edu [223]. The PBS assignation to the correspondent human tRNA type was made by similarity analysis with respect to a tRNA library built from the Transfer RNA database (tRNAdb) of Leipzig University [224] and from a PBS library kindly provided by Professor Jonas Blomberg [8].

2.7 Genomic context of integration

The host genomic context surrounding each HERV-W sequence was characterized by the integration of the HERV-W element genomic coordinates with the UCSC Genome Browser Genes and Genes prediction tracks [225–227]. The elements resulted co-localized with human genes were further analyzed by BLAT search after the activation of OMIM, UCSC, RefSeq and Gencode genes annotations [214].

The presence of predicted Transcription Factors (TFs) putative binding sites was assessed by the integration of HERV-W sequences genomic coordinates with the UCSC Genome Browser Regulation Encode Txn Factor ChIp tracks [228, 229]. TFs binding sites prediction was considered as reliable in the presence of a score ranging from 800 to 1000 (maximum = 1000).

2.8 ORF prediction and putative proteins (puteins) analysis

Amino acid sequences of puteins were obtained from the reconstructions of retroviral Gag, Pol and Env proteins. In particular, ORFs and sequences of puteins were predicted and further confirmed by using i) ReTe online version (http://www2.neuro.uu.se/fysiologi/jbgs/) for ORF prediction and reconstruction [230], ii) Geneious platform [217] ORF prediction tool and iii) further six-frame translations of each gene region. Obtained putein sequences were aligned and further compared with known proteins consensus sequences or other sequences of interest.

Regarding HERV-W and MSRV Env puteins, the obtained full-length and nearly full-length puteins were aligned with respect to ERVWE1/Syncytin-1 precursor (NCBI reference sequence NP_055405.3) to annotate all frameshifts and stop codons and characterize the conservation of domains relevant for Syncytin-1 *in vivo* biological functions.

2.9 Analysis of MSRV sequences

Previously published MSRV sequences and probes were retrieved from GenBank and analyzed by BLAT search for the best matching HERV-W locus/loci based on nucleotide sequence similarity in GRCh37/hg19 assembly. Alignments of MSRV sequences and the relative best matching HERV-W elements were manually inspected on Geneious platform, and discordant positions were annotated. The HERV-W locus/loci identity was then confirmed through the software Recco [231] as previously described [151] with respect to our whole HERV-W dataset.

Chapter 3. Contribution of type W human endogenous retrovirus to the human genome: characterization of HERV-W proviral insertions and processed pseudogenes

3.1 Introduction

As reported in Chapter 1, the human genome harbors an impressive proportion of retroviral endogenous sequences, accounting for its 8%, whose persistence during evolution led to the accumulation of several mutations, insertions and deletions, that have generally compromised their coding capacity [232]. Among HERVs, the HERV-W group is a prominent exception. Initially identified for its possible role in Multiple Sclerosis (MS), this group showed a high and specific expression level in placental tissues. Further investigations interestingly revealed that an HERV-W provirus, named ERVWE1 and localized to locus 7q21.2, i) retained a complete *env* ORF [23]; ii) was able to produce a functional protein, called Syncytin-1 and iii) had been co-opted by the human genome for the trophoblast cells fusion during pregnancy, an important structure for regulating the exchange between mother and fetus [24, 25, 48].

Starting from these findings, as extensively described in the introduction section, the general expression of HERV-W group has been investigated in the different tissues, above all to find a correlation to various human diseases, such as MS and other neurological and neuropsychiatric disorders, as well as various cancers and autoimmune diseases. However, despite the great interest in HERV-W expression, no definitive correlation with human pathologies have been conclusively demonstrated so far [100] and the characterization of the group at the genomic level still remained a major genetic goal and a bioinformatics challenge [233]. Specifically, the knowledge of the HERV-W genomic distribution and number of copies is still referred to analyses performed a few years ago, using different methodologies ([33, 34, 36]). In particular, Voisset et al. (2000) described the presence of 70, 100, and 30 HERV-W-related *gag*, *pro*, and *env* regions respectively, using a PCR approach with HERV-W genes specific primers on isolated chromosomes [36]. Costas (2002) identified a total of 140 HERV-W elements through a NCBI BLAST within the draft sequence of the human genome [34]. Pavlícek et al. (2002) reported 311 HERV-W elements and 343 solitary LTRs identified using the RepeatMasker program in the GoldenPath assembly of 87% of the human genome [33]. These works obviously represent milestones in the HERV-W group characterization, but the absence of a complete and exhaustive version of the human genome at the time and the use of different methodologies make it hard to compare and correlate these data with the current version of the human genome, and to the expression profiles obtained in physiological and pathological contexts. Moreover, with the exception of the well-described Syncytin-1 provirus, detailed information about the group composition and its members characteristics are somehow lacking, preventing until now a comprehensive analysis of their possible involvement in human pathologies. In fact, a detailed knowledge of HERV-W transcripts genomic origin is essential to properly evaluate the great amount of expression data, and to evaluate specific HERV-W elements possible involvement in disease development and/or progression. Furthermore, it is well known that the mere presence of HERV integrated elements could affect human physiology and health through alternative mechanisms even in the absence of gene expression or products. This can occur for example i) with gene physical disruption after HERV insertion [234, 235]; ii) by damaging recombination events that can produce genomic alterations ranging from deletions and duplications to large-scale chromosomal rearrangements [236]; and iii) through the effects exerted by HERVs and their LTRs that naturally present promoters, enhancers, polyadenylation signals and splice donor sites [8, 237–239] and can regulate also human gene expression in a tissue specific manner [240–250].

In this context, the current HERV-W expression studies seem to be not exhaustive to understand the real effects that these elements can exert. In fact, on the one side, due to their multi-copy nature, it is not always clear from which genomic locus a HERV-W mRNA is transcribed, and, on the other side, the potential effects of such sequences is not solely connected to their expression capacity, but depends also on their localization and their ability to (dys)regulate host functions also through alternative mechanisms behind the presence of a RNA/protein products.

In the light of this, the definition of a precise and updated HERV-W genomic map is a pressing need to better evaluate their role in human health and their real influence on host genome.

3.2 HERV-W identification and general classification

In a recent work aimed to the global classification of HERV clades and sequences in the human genome, we reported the presence of 126 elements belonging to HERV-W group [8]. These data were obtained through the bioinformatics ReTe tool, a program package implemented for the identification of ERV full integrations in vertebrate genomes and the attempted reconstruction of the relative ORFs and proteins [212]. For HERV sequences recognition ReTe uses a collection of generic, conserved motifs, a few within *env* and *gag* genes, that can be mutated or lost in defective proviruses [8]. Such "bias" was reported as responsible for the low representation of HERV Class III proviruses that have an aberrant *gag* and may not have an *env* [8]. In the light of this, willing to build an updated dataset of HERV-W sequences in the human genome GRCh37/hg19 assembly, we used a combined strategy based on i) the ReTe analysis and ii) a traditional Genome Browser BLAT search [213], using the assembled RepBase reference LTR17-HERV17-LTR17 [251] as a query. This integrated approach led to the characterization of a total of 213 HERV-W related sequences: the 126 previously identified by ReTe and 87 additional elements retrieved by Genome Browser BLAT. Indeed, a high proportion of newly identified HERV-W sequences showed huge and recurrent deletions that caused loss of extended portions in *gag*, *pol* and *env* genes (described more in detail in the structural characterization section). Hence, the defective nature of the great majority of HERV-W sequences could be responsible for the underrepresentation of the ReTe outcome, confirming the importance of a double approach in HERV identification.

The main characteristics of HERV-W elements are summarized in Table 6.

Table 6. *HERV-W elements identification in Human Genome assembly GRch37/hg19*

Locus	Strand	Start	End	Type	Sg[a]	PBS type	O.C.A[b]	Age[c]
1p34.2	+	42410127	42415982	pseudogene	1	-[d]	Rhesus	32,52
1p33a	-	46851453	46856995	pseudogene	1	W	Gibbon	26,99
1p33b	+	47417563	47424579	pseudogene	1	W	Rhesus	33,82
1p32.3a	+	51692797	51696169	pseudogene	2	-	Rhesus	30,11
1p32.3b	-	55376682	55385198	provirus	1	W	Rhesus*	26,82
1p32.2	-	56248856	56254641	provirus	2A	F	Rhesus	32,99
1p22.2a	+	89390212	89397564	pseudogene	1	W	Rhesus	32,09
1p22.2b	+	91644689	91646698	pseudogene	1	-	Rhesus	24,73
1p13.3	-	110394855	110400764	provirus	1	-	Orangutan	29,30
1p12	-	119710922	119713987	pseudogene	1	-	Rhesus	26,34
1q22	-	155592180	155596455	undefined	-	-	Orangutan	-
1q25.2	-	178110353	178112194	pseudogene	2	-	Rhesus	39,24
1q32.1	-	205835460	205840608	pseudogene	2	-	Rhesus	36,76
1q32.3a	-	212029138	212031215	pseudogene	1	-	Rhesus	31,32
1q32.3b	+	212031216	212032263	undefined	-	-	Rhesus	-
1q42.13	-	227812215	227821260	provirus	2A	F	Rhesus	39,88
2p25.3	-	153394	160782	provirus	1	W	Rhesus*	24,45
2p24.2	+	17520208	17527981	provirus	1	R	Gibbon	23,94
2p23.1a	+	30738819	30743130	provirus	2A	F	Rhesus	39,00
2p23.1b	+	31854223	31859227	pseudogene	1	R	Orangutan	35,12
2p22.3	+	33882489	33889522	pseudogene	1	R	Rhesus	35,90
2p16.2	+	53983776	53988563	pseudogene	2	W	Rhesus	39,45
2p12a	+	76098816	76106624	provirus	1	W	Gibbon	30,23
2p12b	-	79390297	79397477	pseudogene	1	R	Rhesus	25,22
2q11.2	-	96882882	96889063	pseudogene	2	W	Gibbon	31,15
2q12.2	+	106825670	106832132	pseudogene	1	W	Rhesus	19,27
2q13	-	112796923	112802566	pseudogene	1	W	Rhesus	35,18
2q22.1	-	136959456	136960929	pseudogene	1	-	Gibbon	40,75
2q22.2	-	143656248	143665468	provirus	2B	N	Rhesus	45,10
2q22.3	+	147756109	147758691	undefined	-	R	Rhesus	-
2q24.3	-	165514421	165516121	pseudogene	1	-	Gibbon	23,59
2q31.1a	-	172416543	172418614	pseudogene	1	-	Gibbon	49,74
2q31.1b	+	176189870	176191257	pseudogene	1	-	Gibbon	39,76
2q31.2a	-	178273969	178277886	pseudogene	2	W	Rhesus	39,66
2q31.2b	+	179327137	179328271	pseudogene	1	-	Gibbon	36,14
2q31.3	+	181096404	181104046	provirus	2A	I	Rhesus	42,88
2q32.3	-	196199400	196204796	pseudogene	2	P	Rhesus	38,56
2q35	-	218455569	218457491	pseudogene	1	-	Rhesus	31,16
2q37.3	-	239902211	239904111	pseudogene	1	-	Rhesus	24,56
3p24.3	-	19913287	19918128	pseudogene	1	W	Gibbon	24,16
3p24.1	+	27042514	27047921	provirus	1	W	Gibbon	28,64
3p22.2	-	38330909	38338816	provirus	1	R	Rhesus	30,01
3p22.1	+	39699867	39701564	pseudogene	1	-	Rhesus	22,10
3p21.31	-	48372016	48377731	pseudogene	2	?[e]	Rhesus	42,89

3p12.3	-	74921984	74927237	pseudogene	1	R	Rhesus	35,77
3p11.1	-	87989983	87995529	pseudogene	1	W	Gibbon	25,93
3q11.2	-	96385297	96394288	provirus	1	W	Gibbon	28,33
3q13.31	+	115463422	115469539	pseudogene	2	-	Rhesus	35,56
3q13.32	+	118561539	118570538	undefined	-	-	Rhesus	-
3q22.1	-	130915705	130918941	pseudogene	1	-	Orangutan	22,17
3q22.2	+	135654191	135657750	pseudogene	1	-	Gibbon	21,10
3q23a	+	141538627	141540355	pseudogene	1	-	Orangutan	20,93
3q23b	+	142157021	142162495	pseudogene	1	-	Gibbon	25,05
3q25.1a	-	149474249	149476612	pseudogene	1	-	Rhesus	27,41
3q25.1b	+	150572043	150579124	pseudogene	1	W	Rhesus	21,75
3q25.2	+	154695864	154697090	pseudogene	1	-	Gibbon	59,38
3q26.1a	+	162040467	162046157	pseudogene	1	-	Gibbon	27,84
3q26.1b	-	163412265	163418737	provirus	2A	-	Rhesus	51,85
3q26.31	+	172438594	172441229	undefined	-	-	Rhesus	-
3q26.32	+	178772379	178777943	pseudogene	1	W	Rhesus	27,61
3q28	+	191376573	191383381	pseudogene	1	W	Gibbon	25,48
4p16.3	-	185552	190863	provirus	2A	W	Rhesus	36,18
4p16.1	+	8422092	8429492	provirus	2A	?	Rhesus	46,40
4p15.1	+	33743127	33744972	provirus	1	-	Gibbon	52,41
4p14	-	36442153	36445845	pseudogene	2	-	Rhesus	39,49
4p13	-	42287455	42294913	provirus	1	W	Rhesus	29,50
4q13.1	+	63836735	63843478	provirus	2A	E	Rhesus	46,58
4q13.3	-	73791293	73798903	provirus	1	W	Gibbon	28,59
4q21.22	+	83394470	83401206	provirus	1	R	Rhesus	35,60
4q21.23	-	86273279	86276547	pseudogene	1	W	Rhesus	28,85
4q23	-	100025890	100031105	pseudogene	2	F	Gibbon	23,95
4q24	+	106408946	106410359	undefined	-	-	Rhesus	-
4q25	-	111200009	111205600	pseudogene	1	W	Rhesus	28,28
4q26	+	114965536	114972972	pseudogene	1	W	Gibbon	26,84
4q28.3	+	133803451	133810295	pseudogene	2	W	Rhesus	42,06
4q31.1	-	139542941	139548353	provirus	1	-	Rhesus	32,78
4q31.3	+	153762980	153763482	pseudogene	1	-	Rhesus	34,79
4q32.3	+	165576784	165579304	pseudogene	1	-	Rhesus	46,94
4q33	-	171111343	171118255	pseudogene	2	W	Rhesus	37,72
4q35.1	+	183920546	183925533	pseudogene	1	-	Gibbon	18,01
5p13.3	-	31395287	31397921	undefined	-	-	Gibbon	-
5p13.2	-	36328801	36333224	pseudogene	1	R	Rhesus	30,76
5p12	+	44111190	44115179	provirus	1	W	Gibbon	22,10
5q11.2	-	56815850	56818716	pseudogene	1	-	Rhesus	27,16
5q12.1	+	59954322	59962288	provirus	2B	I	Rhesus*	28,76
5q14.3a	-	87472587	87473700	pseudogene	1	-	Gibbon	23,30
5q14.3b	+	89089703	89090854	provirus	1	-	Gibbon	62,97
5q21.3	-	107909807	107914283	provirus	1	-	Gibbon*	25,28
5q22.2	+	111779389	111787195	provirus	1	R	Orangutan*	20,46
6p25.3	-	1330281	1331849	pseudogene	1	-	Orangutan	36,94
6p23	+	13878567	13883955	pseudogene	2	P	Rhesus	32,91
6p22.3	-	24676916	24683328	provirus	1	W	Gibbon	23,31

6p12.2	-	52779674	52786712	provirus	1	R	Rhesus*	37,19
6q12	+	65714885	65721860	pseudogene	1	W	Gibbon	29,41
6q14.1	+	82046663	82053574	provirus	1	W	Gibbon*	34,15
6q14.2	+	84153479	84160109	pseudogene	1	W	Rhesus	40,08
6q14.3a	-	85421420	85430868	provirus	2A	N	Rhesus	35,48
6q14.3b	+	85653935	85655028	pseudogene	1	-	Orangutan	29,31
6q15	-	89123912	89131226	pseudogene	1	W	Gibbon	24,15
6q21a	+	106676012	106683689	pseudogene	1	R	Rhesus	29,31
6q21b	+	107620616	107621947	undefined	-	-	Rhesus	-
6q21c	+	111452221	111459275	pseudogene	2	S	Rhesus	43,50
6q23.3	-	138040065	138043839	pseudogene	1	-	Rhesus*	32,97
6q24.2a	-	143400938	143408846	pseudogene	2	F	Rhesus	30,65
6q24.2b	-	144437462	144442702	pseudogene	1	W	Rhesus	24,99
6q27a	+	166727385	166729870	undefined	-	-	Rhesus*	-
6q27b	+	167304571	167311741	pseudogene	1	W	Rhesus	30,55
7p21.3	-	12892571	12893302	pseudogene	2	-	Rhesus	42,64
7p21.1	-	16716694	16717518	pseudogene	1	-	Rhesus	39,87
7p14.2	+	35735748	35736627	provirus	1	-	Rhesus	60,98
7p14.1	-	40207037	40213362	provirus	2B	R	Rhesus	33,16
7q21.2	-	92097313	92107506	provirus	1	W	Rhesus	22,56
7q31.1a	+	107977725	107984934	pseudogene	1	W	Rhesus	27,33
7q31.1b	-	114019143	114026368	pseudogene	1	W	Gibbon	29,70
7q31.31	+	119206200	119210883	pseudogene	1	-	Gibbon	29,07
7q31.32	+	121811709	121817649	provirus	2B	-	Rhesus	49,00
7q32.3	-	131505601	131509683	provirus	2A	-	Rhesus	34,32
7q33	+	134270175	134277767	provirus	2A	F	Rhesus	31,40
7q36.1	+	149368691	149374312	provirus	2B	?	Rhesus	37,57
8p21.3	+	20012257	20017145	pseudogene	2	F	Rhesus	45,74
8q11.21	-	49148502	49151947	pseudogene	1	-	Orangutan	18,77
8q12.1	-	61331973	61338078	pseudogene	2	I	Rhesus	27,03
8q12.3a	-	63507960	63509929	pseudogene	1	-	Gibbon*	17,69
8q12.3b	+	65675513	65680917	pseudogene	2	L	Rhesus	31,16
8q13.2	+	68834197	68835866	pseudogene	1	-	Rhesus	16,94
8q21.11	+	74734099	74734838	pseudogene	1	-	Gibbon	34,35
8q21.13	+	81651951	81655427	pseudogene	1	-	Rhesus	28,96
8q24.13	-	125912007	125919468	provirus	2A	?	Rhesus	32,19
9p24.1	+	8747223	8748523	pseudogene	1	-	Gibbon	25,00
9p21.3	+	22823028	22823838	pseudogene	1	-	Rhesus	60,62
9p21.1	-	29656373	29658925	undefined	-	-	Rhesus	-
9p13.3	-	35640305	35642827	provirus	1	-	Gibbon	25,06
9q22.1	-	91554926	91559339	pseudogene	1	R	Rhesus	29,73
9q22.31	+	94742841	94744363	pseudogene	1	-	Rhesus	40,89
9q31.3	+	114098689	114100459	pseudogene	1	-	Gibbon	20,01
10p12.2	-	23585121	23591090	pseudogene	1	R	Gibbon	28,02
10q11.22	-	49873340	49875049	pseudogene	1	-	Gibbon	27,58
10q21.2	-	62793461	62799671	provirus	2A	W	Rhesus	31,53
10q21.3	+	65804516	65805376	pseudogene	1	-	Gibbon	32,31
10q23.1	-	86285127	86290825	provirus	1	R	Rhesus	32,39

10q23.33	-	96594327	96601288	pseudogene	1	R	Gibbon	28,63
10q24.1	-	97477116	97484627	pseudogene	2	W	Rhesus	29,80
11p15.4	-	9369352	9371164	pseudogene	1	-	Rhesus	34,47
11p14.3a	-	22334363	22339865	provirus	1	W	Gibbon	31,04
11p14.3b	+	26023135	26027273	provirus	-	-	Rhesus	32,28
11p14.2	-	26611978	26619222	provirus	2B	G	Rhesus	29,72
11p12	+	38623139	38628790	provirus	2A	F	Rhesus	41,84
11q14.1	-	77569373	77574526	pseudogene	1	W	Rhesus	27,24
11q14.2	-	86544534	86547860	pseudogene	1	-	Rhesus	29,71
11q22.3	-	107851810	107853843	pseudogene	1	-	Gibbon	39,44
11q23.3	-	117907500	117908549	pseudogene	1	-	Rhesus	43,37
12p13.31a	+	7333743	7339129	pseudogene	2	F	Rhesus	37,45
12p13.31b	-	8914006	8921050	pseudogene	2	W	Rhesus	34,30
12p11.1	+	34253056	34255891	pseudogene	1	-	Gorilla	25,35
12q12a	+	38865383	38868227	pseudogene	1	-	Gibbon	23,33
12q12b	+	39030303	39035353	pseudogene	2	?	Gibbon	28,41
12q12c	+	40777438	40783340	provirus	2A	?	Rhesus	34,25
12q13.12	-	51296259	51307150	provirus	1	W	Rhesus*	34,11
12q13.3	-	57360856	57362378	undefined	-	-	Homo	-
12q14.1	-	59245647	59253534	provirus	1	W	Gibbon	23,29
12q21.31	+	85187760	85192453	pseudogene	2	?	Rhesus	28,96
12q23.3	+	105337032	105337822	pseudogene	1	-	Rhesus	36,98
12q24.31	+	124032964	124043310	provirus	2A	I	Rhesus	33,77
12q24.33	-	132357772	132365497	provirus	1	R	Rhesus	35,32
13q13.3	+	37530787	37533118	pseudogene	2	-	Rhesus	34,09
13q21.1	+	55627766	55635877	provirus	1	R	Gibbon	19,09
13q21.31	-	65279823	65282432	pseudogene	1	-	Rhesus	41,48
13q21.33	+	69795752	69799468	provirus	1	W	Gibbon	39,06
13q31.1	+	83031689	83034708	pseudogene	1	W	Gibbon	25,74
13q31.3	+	93693066	93695685	undefined	-	-	Rhesus	-
14q11.2	+	22704447	22712069	provirus	1	S	Gibbon*	24,00
14q12	+	26728928	26730215	provirus	1	-	Gibbon	32,55
14q21.2	-	45488688	45492897	provirus	-	-	Rhesus	36,49
14q22.1	+	53828360	53829723	pseudogene	1	-	Gibbon	22,61
14q23.1	+	58588999	58592862	pseudogene	1	-	Rhesus	33,96
14q32.11	-	91692008	91693212	pseudogene	1	-	Gibbon	22,43
15q21.3	-	55597080	55604574	pseudogene	1	W	Rhesus	26,32
15q22.32	-	67261563	67268962	pseudogene	1	R	Rhesus	36,02
15q26.1	+	92841377	92845430	pseudogene	1	W	Rhesus	40,53
17q12a	+	33878307	33879612	pseudogene	1	-	Orangutan	23,10
17q12b	+	35689888	35693622	pseudogene	1	-	Rhesus	35,26
17q21.33	-	48476362	48483068	provirus	1	W	Rhesus	33,89
17q22	-	53088886	53095859	pseudogene	1	W	Rhesus	41,51
18p11.31	+	4681680	4692409	provirus	1	S	Orangutan	31,11
18p11.21	+	13656678	13657837	pseudogene	1	-	Rhesus	21,27
18q21.32	+	58377841	58380952	pseudogene	1	-	Gibbon	17,21
18q21.33	+	60149574	60150323	pseudogene	1	-	Gibbon	25,04

19p12a	+	20210345	20212271	provirus	2	-	Gorilla	58,42
19p12b	+	21812187	21817584	pseudogene	1	W	Gorilla	22,90
19p12c	+	22928203	22934876	pseudogene	1	W	Gibbon	25,59
19p12d	+	24129884	24130952	pseudogene	1	-	Rhesus	29,50
19q13.2a	+	40577847	40578561	pseudogene	1	-	Rhesus	36,86
19q13.2b	-	41492147	41494054	provirus	2	-	Rhesus	62,29
20p13	-	171333	174808	pseudogene	1	-	Gibbon	26,76
20q13.2	+	53965519	53969590	pseudogene	1	-	Gibbon	30,15
21q21.1	-	20125060	20132866	provirus	1	W	Gibbon	24,54
21q21.3	+	28226756	28234297	provirus	1	W	Rhesus	27,93
21q22.2	-	41111191	41118353	pseudogene	1	W	Gibbon	27,40
22q12.3	-	34345530	34345988	pseudogene	1	-	Gibbon	38,21
Xp22.31	-	7617210	7622109	pseudogene	1	W	Rhesus	28,65
Xp11.3	+	46046300	46049013	pseudogene	2	-	Rhesus	36,42
Xp11.21	-	57424992	57432088	pseudogene	1	W	Rhesus	39,21
Xq12	-	65517016	65519281	pseudogene	1	-	Rhesus	33,82
Xq13.3	-	75118132	75124909	provirus	1	W	Gibbon	25,72
Xq21.1a	-	77602103	77606013	pseudogene	1	W	Gibbon	43,79
Xq21.1b	+	79211390	79218603	pseudogene	1	R	Orangutan	16,68
Xq22.3a	+	105244784	105248113	pseudogene	1	-	Rhesus	25,84
Xq22.3b	-	106295361	106298094	pseudogene	1	-	Rhesus	18,64
Xq23	-	115,871,924	115,875,415	pseudogene	1	-	Gibbon	22,36
Xq26.2	-	131063277	131067083	undefined	-	-	Rhesus	-
Xq27.1	+	139159755	139162980	pseudogene	1	-	Rhesus	23,46
Yp11.2	-	7754065	7760683	provirus	2B	?	Chimpanzee	72,52
Yq11.222	+	21241822	21249383	provirus	1	W	Chimpanzee	55,06

[a] HERV-W subgroup based on phylogenetic analysis of 5′ and 3′ LTRs and gag and pol genes, supported by the identification of key mutations with respect to the RepBase Update generated reference LTR17-HERV17-LTR17

[b] Oldest common ancestor, refers to most distant species that shares the sequence. Identified by RepeatMasker HERV17 annotations after BLAT searching for the primates orthologous locations in the Chimpanzee, Gorilla, Orangutan, Gibbon, Rhesus and Marmoset available genomes on the UCSC Genome Browser

[c] estimated time of integration, in million years

[d] Impossible to define due to the lack of the retroviral portion involved in classification

[e] Impossible to unambiguously assign

*Sequence found as solitary LTR

We named individual HERV-W elements according to their genomic localization, in order to have a unique and direct identification of each sequence. In the presence of multiple sequences in the same locus, the order within the band is expressed with a letter following the alphabetical order, as previously described [252]. HERV-W elements occurred on all chromosomes showing no recognizable cluster distribution, except chromosome 16 that apparently do not contain HERV-W proviruses or processed pseudogenes.

The 213 HERV-W sequences were firstly divided into three categories due to previously reported structural characteristics that mostly address the LTRs portion and that reflect their mechanism of formation [33]: proviruses (65), processed pseudogenes (135) and undefined elements (13). Briefly, with respect to the LTR17 RepBase consensus (780 nucleotides), proviral sequences show complete LTRs (referred here as proviral LTRs) and have been inserted into human DNA by a traditional process of retroviral integration. Proviral LTRs show a traditional composition with two unique regions (U3 and U5) separated by a repeated portion (R), giving a U3-R-U5 structure. As described by Pavlícek et al. [33], processed pseudogenes are LINE-1-retrotransposed HERV-W sequences presenting i) truncated LTRs (referred here as pseudogenic LTRs), with the 5′LTR showing a R-U5 structure (start from nucleotide 256 of the consensus) and the 3′LTR showing a U3-R structure (end at position 326 of the consensus), ii) a poly(A) tail of variable length, and iii) a common TT/AAAA insertion motif and a variable-length (5–15 bp) target site duplication [33]. Finally, undefined elements are sequences that have lost those regions in both LTRs and so remained undefined due to the absence of the signatures described above.

It is interesting to note that our results differed from previous analysis performed a number of years ago on not exhaustive draft versions of the human genome [33, 34, 36] and with the use of different detecting methodologies, leading to discordant results that are not always easy to retrieve and correlate with current data. In fact, on one side, two studies on HERV-W distribution and composition [34, 36] reported a lower number of elements with respect to our dataset. In particular, Voisset et al. described the presence of 70, 100, and 30 HERV-W-related *gag*, *pro*, and *env* regions, respectively, without further indications about their origin [36], while Costas identified a total of 140 HERV-W elements, 73 less than the present analysis. On the other side, when compared to our dataset, the study by Pavlicek et al. reported a higher number of HERV-W sequences (311) [33]. The lack of available supplementary information of Pavlicek HERV-W dataset (e.g. nucleotide sequences or genomic localization) did not allow us to perform a direct comparison with our results. However, Pavlicek et al. HERV-W sequences were retrieved from a draft assembly version of the 87% of the human genome using the RepeatMasker program that, in the presence of the recurrent and huge deletions such as the ones observed in the HERV-W sequences, could not easily identify the whole elements. Hence, more fragments previously reported as independent elements possibly belonged to the same provirus/pseudogene. This hypothesis seems to be confirmed by a subsequent study where the same dataset has been used for the HERV-W processed pseudogenes length distribution analysis [253]. Such report showed that the most represented length class in Pavlicek dataset enclosed very short sequences (0-0,5 Kb), with a low proportion of > 3,5 Kb elements. Differently, in our dataset > 90% of sequences are in the 1-7,5 Kb range, with around 25% > 6,5 Kb (data not shown). Overall, the use of the Rete software, that relates retroviral elements reconstructing the original chain [7], together with a visual inspection of all aligned sequences plus their flanking sites of integration with respect to the group reference, probably led to more reliable sequence recognition. Furthermore, the overestimation of HERV-W members in Pavlicek dataset could also be due to the possible inclusion of HERV9 sequences, highly related to HERV-W but constituting a separate phylogenetic group [8]. In fact, to avoid such bias we initially included a HERV9 consensus in every HERV-W phylogenetic trees, assuring that none of the sequences classified as HERV-W clustered with HERV9 group (data not shown). Importantly, a significant contribution on the HERV-W group presence in the human genome was recently provided in a study in which the cDNA obtained from HERV-W RNA transcripts in MS patients and controls brain samples was amplified in the *env* region and assigned to single HERV-W loci by Genome Browser BLAT on the NCBI36/hg18 assembly (March 2006) [119]. Although the purpose of that study was not a HERV-W group genomic characterization and methodology was biased for *env* sequences analysis, yet it provided a remarkable genomic map of 176 HERV-W loci, enclosing 35 proviruses, in their supplementary material [119]. Noteworthy, with respect to this study, our analysis led to the identification of 37 further HERV-W elements (9 proviruses, 18 processed pseudogenes and 10 undefined sequences), and to a more defined classification of proviruses and processed pseudogenes.

3.3 Structural characterization

In order to characterize the HERV-W structure, we firstly aligned and analyzed the 213 retrieved HERV-W sequences with respect to the assembled reference LTR17-HERV17-LTR17, built from RepBase Update consensus sequences for HERV-W LTRs and internal portion [251]. HERV-W sequences showed a typical proviral structure (Figure 5), with the *gag* (nucleotides 2718-4191), *pro* (4195-4449), *pol* (4450-7692) and *env* (7720-9348) genes flanked by 5′ and 3′ LTRs (1-780 and 9406-10186, respectively) (start and end positions are referred to LTR17-HERV17-LTR17). In addition, almost all HERV-W identified sequences present a 2 Kb long non-coding region, that we named *pre gag,* located between 5′LTR and *gag* gene and characterized by an AG rich expansion of variable length. This portion was previously reported for three cDNA HERV-W clones [23], but neither function or origin has been proposed or demonstrated yet.

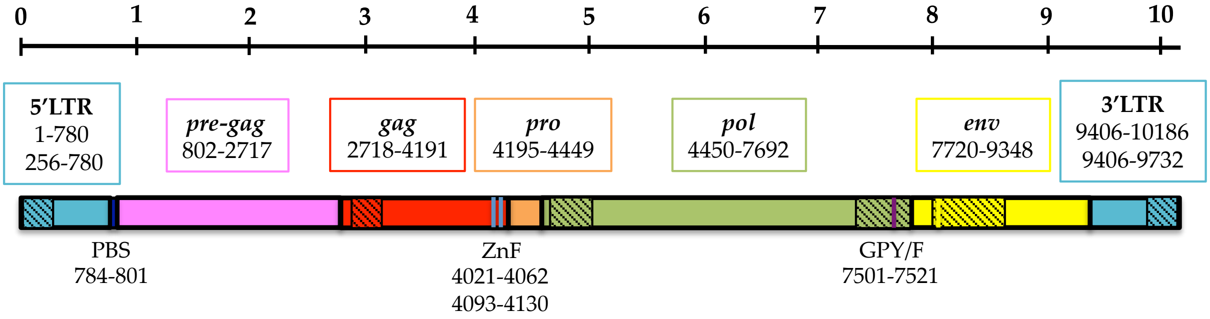


***Figure 5**. HERV-W sequences general structure.*
gag, pro, pol and *env* genes start and end nucleotide positions, as well as the coordinates for 5′ and 3′LTRs for both proviruses and processed pseudogenes, are annotated in the group assembled RepBase consensus, LTR17-HERV17-LTR17. Also the group main structural features, i.e PBS, Gag NC Zinc Fingers (ZnF) and Pol IN GPY/F motifs, are reported. The genes portions with black lines correspond to regions associated to frequent deletions, i.e. recurrent in a high percentage of sequences: *gag* nucleotides 2780-3209 (45%); *pol,* nucleotides 4513-6184 and 6797-7692 (28% and 84%, respectively); and *env,* nucleotides 7928-9114 (85%), with the exception of a small region at position 8289-8318.

Firstly, comparing proviral versus pseudogenic sequences, we asked whether, beside the shorter pseudogenic LTRs, the internal sequence completeness was comparable. We evaluated the presence of each retroviral element (5′LTR, *gag*, *pro*, *pol*, *env* and 3′LTR) in the three classes of HERV-W sequences (proviruses, processed pseudogenes and undefined), considering as retained all elements with at least the 20% of nucleotides with respect to the correspondent element in LTR17 and HERV17 RepBase consensus (Figure 6).

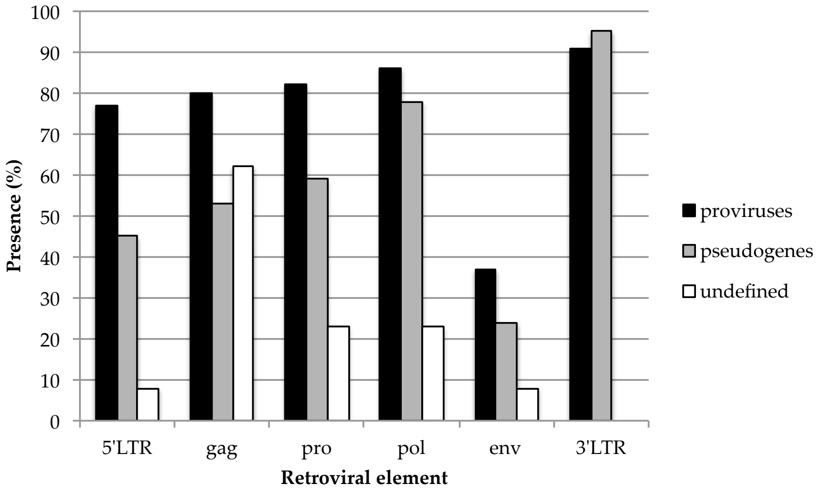


***Figure 6.** Overview on the HERV-W structural completeness.*
The presence of each retroviral element (LTRs and genes) in the three classes of HERV-W sequences is depicted. An element is considered as retained if at least the 20% of its nucleotide sequence is present (lengths are referred to LTR17 and HERV17 reference elements). With respect to the LTR17 RepBase consensus (780 nucleotides), proviral sequences show complete LTRs, while processed pseudogenes present typically truncated LTRs due to

the L1-mediated retroposition, lacking U3 in 5′LTR (R-U5 structure, nucleotides 256-780) and U5 in 3′LTR (U3-R structure, nucleotides 1-326).

Analysis of all identified HERV-Ws showed that proviral sequences appear to be the most complete, with a better general maintenance of the considered retroviral structures, while pseudogenes, interestingly, frequently lack the 5′LTR, absent in > 50% of elements, and *gag* and *pro* genes. Importantly, for all classes, *env* is the most frequently lost element, due to recurrent extended deletions that involve > 80% of the gene. In addition, besides the lack of both LTRs, undefined sequences showed a higher frequency of gene loss, especially in the *pro-pol-env* portions, indicating that deletion processes were not limited to LTR sequences.

Secondly, in order to define the group and the single sequences structural characteristics, the 213 HERV-W elements were further analyzed in great detail by annotating all insertions/deletions with respect to the consensus LTR17-HERV17-LTR17, as schematically represented for the 59 proviruses with minimum length of 2,5 Kb in Figure 7.

In comparison to the consensus, in all types of sequences some recurrent deletions clearly affect viral genes (as shown also in Figure 5), resulting in the loss of some big viral portions: i) nucleotides 2780-3209 in *gag* gene (45% of the sequences), ii) nucleotides 4513-6184 and 6797-7692 (IN portion) in *pol* gene (28% and 84% of the sequences, respectively), and iii) nucleotides 7928-9114 in the *env* gene (85% of the sequences), with the exception of a small region of about 30 nucleotides at position 8289-8318 that is frequently present despite the flanking deletions. Interestingly, the recurrent loss of *pol* and *env* genes, deleted in the C-terminal IN portion and retaining only the TM intracytoplasmic tail, respectively, possibly suggests a selective removal of regions that were no longer needed in the absence of an active infective transmission.

In addition to these major mutations, the analyses highlighted a greater amount of minor insertions/deletions and single nucleotides substitutions that, overall, allow to specifically identify the uniqueness of each HERW-W sequence.

The majority of these variations appear to be casually distributed among the sequences, as expected from the normal random genomic substitution rate, while a number of them is shared by the great majority of the sequences and characterize their structure with respect to the reference.

This analysis allowed also to better defining a new HERV-W consensus generated from our proviral dataset that we graphically compared with the LTR17-HERV17-LTR17 consensus (Figure 8).

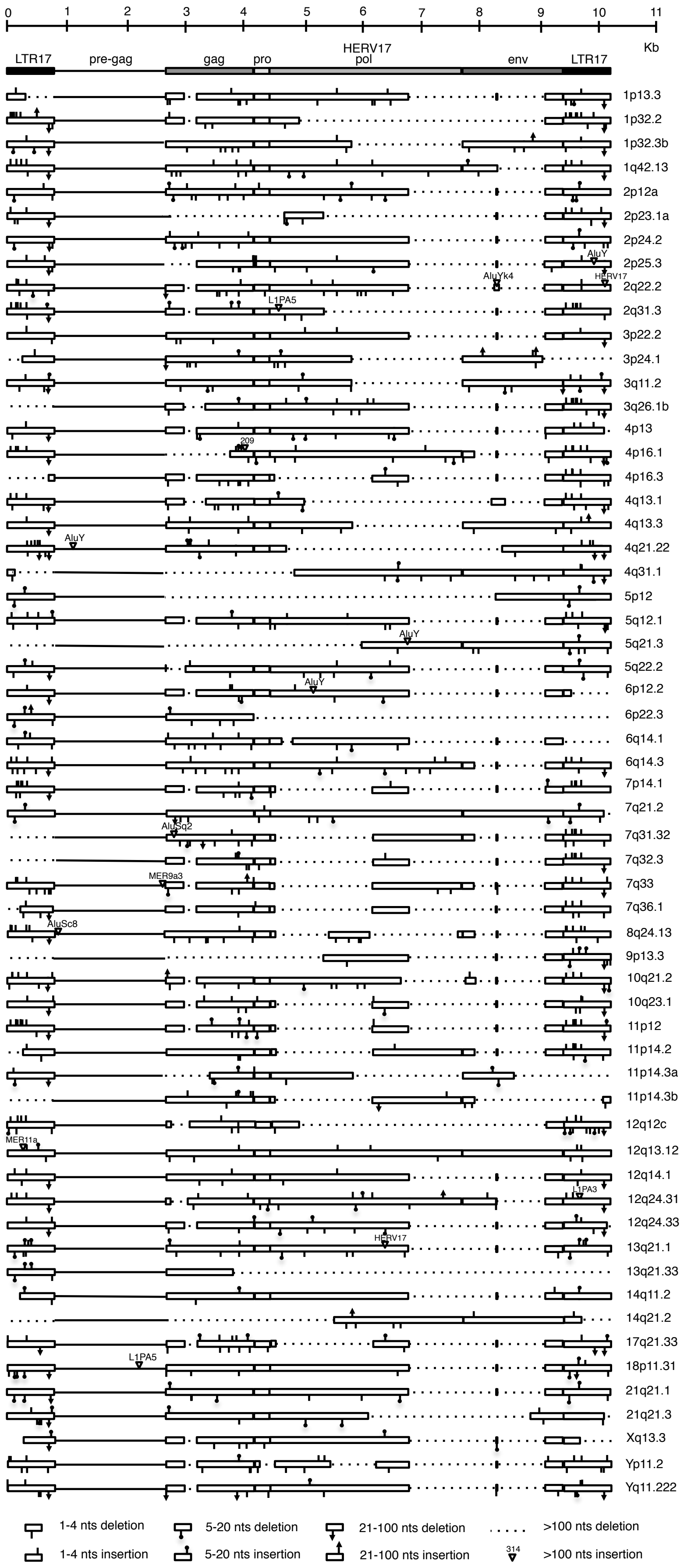


***Figure 7**. Graphical representation of all insertion and deletions ≥ 1 nucleotide in the 59 proviral sequences with length > 2,5 Kb with respect to LTR17-HERV17-LTR17 reference sequence.*

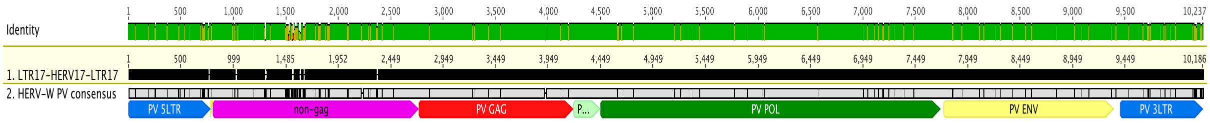


Figure 8. *Comparison between HERV-W RepBase LTR17-HERV17-LTR17 (black) and the proviral dataset generated consensus (grey).*

Nucleotide identity between the two consensus sequences is represented by the colored upper bar (green: 100% identity; greeny-brown: between 100% and 30% identity; red: identity < 30%), while single nucleotide differences of the new consensus with respect to LTR17-HERV17-LTR17 are represented with black lines.

The LTRs and genes localization is shown below.

Interestingly, the LTR structures of the new HERV-W consensus showed recurrent mutations defining two subgroups of sequences that were used, in combination with the phylogenetic analysis, as key positions for subgroup definition.

3.4 Phylogenetic analysis and HERV-W subgroup classification

In order to clarify the phylogenetic and evolutionary relationship within the group, LTRs and viral genes were analyzed through the construction of phylogenetic trees using both NJ and ML methods. Both analysis yielded similar trees, so we included only NJ trees in results. In addition, since HERV9 sequences have been reported to be highly related to HERV-W, to exclude possible misclassifications, a HERV9 generated consensus [8] was initially included in all trees in order to identify any member of this HERV-W related family that could be misinterpreted during the sequence collection. As expected, the HERV9 consensus was clearly separated from the HERV-W sequences, which grouped together showing a 100% bootstrap value in every tree (data not shown).

3.4.1 LTRs phylogenetic analysis

Depending on the presence of full length (proviruses) or truncated (processed pseudogenes) LTRs, proviral sequences 5′ and 3′LTRs were analyzed together in the same phylogenetic tree (Figure 9, panel A); while the truncated structure of pseudogenic LTRs only yields a short region (R; about 90 nucleotides) common to 5′ and 3′LTRs, making necessary a separate analysis (Figure 9, panels B and C, respectively). In all trees, the distribution of proviral and pseudogenic LTRs in two major clusters allowed us to divide them into two distinct subgroups, named 1 and 2. The subgroup of belonging of each HERV-W sequence is reported in Table 6. Within the 213 HERV-W sequences, 69% belong to subgroup 1 (38 proviruses and 108 processed pseudogenes), while 24% of them belong to subgroup 2 (25 proviruses and 27 processed pseudogenes). The remaining 7% was constituted of sequences lacking both LTRs and that, subsequently, could not be classified.

Considering that the subgroup division was generally not well supported by bootstrap values, < 50% except for pseudogenic 3'LTR (90%), the identified HERV-W clusters were further analyzed using common features.

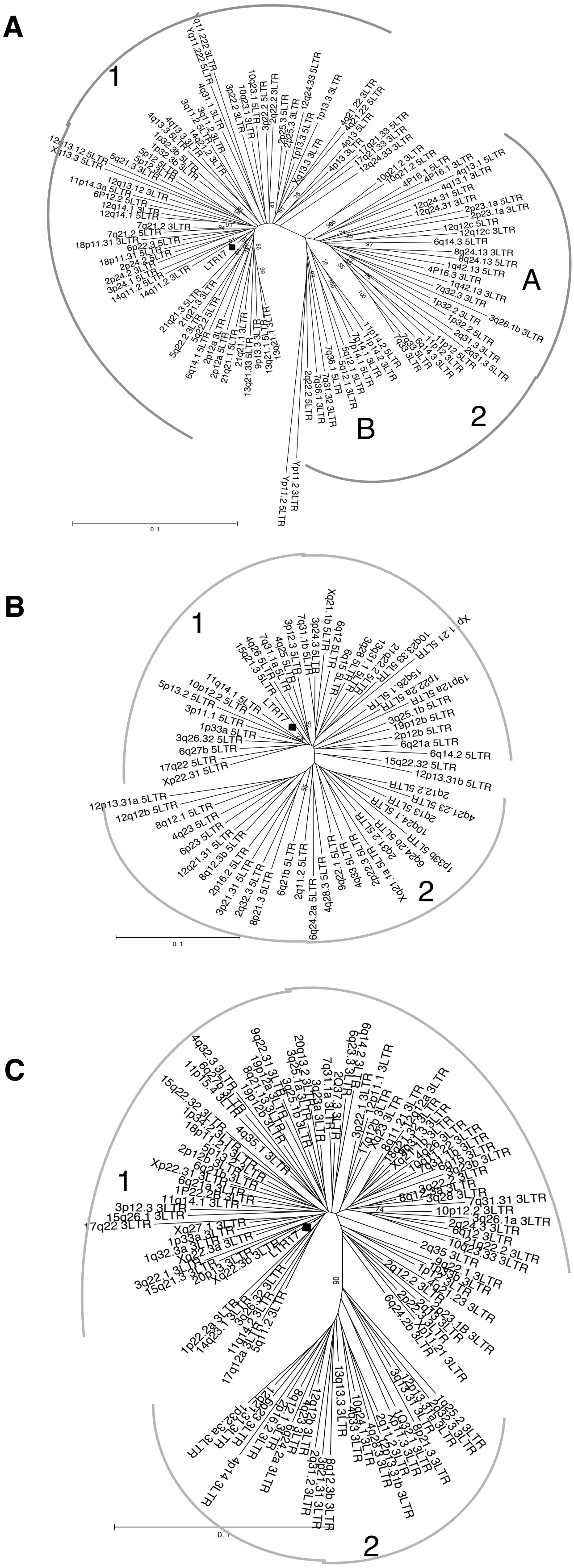


Figure 9. Results from phylogenetic analysis of HERV-W sequences LTRs.

HERV-W proviruses LTRs were analyzed in the same phylogenetic tree (A), while, due to the short region in common, a separate analysis was performed for processed pseudogenes 5′LTRs (B) and 3′LTRs (C). LTR17 RepBase consensus is labeled with a black square. The evolutionary relationships were inferred by using the Neighbor Joining method based on the p-distance model and applying the pairwise deletion option. Phylogenies were tested by using the Bootstrap method with 500 replicates. Length of branches indicates the number of substitutions per site.

3.4.2 LTR-based classification of HERV-W subgroups

The HERV-W sequences belonging to each subgroup (based on their distribution in LTRs phylogenetic analysis) were aligned and compared with respect to LTR17 reference in order to find characteristic features confirming and supporting the robustness of the classification. In general, subgroup 1 elements were not characterized by significant mutations with respect to the reference sequence, probably because LTR17 and HERV17 consensus were built from a few of these elements, such as the 7q21.2 Syncytin-1 locus. A pairwise distance calculation confirmed that the average identity with LTR17 was around 93%. Contrarily, subgroup 2 elements showed a lower identity with respect to LTR17 (87%) and, in fact, the comparison highlighted the presence of recurrent single nucleotide substitutions. The latter were commonly shared with high frequency between subgroup 2 sequences, but not in the other subgroup members, and were thus characterized as key mutations for the proposed classification (Table 7). Particularly, in proviral LTRs we identified 7 positions with characteristic single nucleotide substitutions with respect to LTR17, which are present in 95-100% of subgroup 2 members and rarely found (0-3,5%) among subgroup 1 members. Moreover, in proviral LTRs tree we observed that subgroup 2 elements seemed divided in two further branches, indicated as type 2A (n=16) and 2B (n= 7) (Figure 9, panel A, and Table 7). While both shared the recurrent mutation typical of the main subgroup, each one further shows some additional features found in 90-100% of sequences and rarely present in subgroup 1 elements. These additional mutations were not exclusive of each branch, but were present also in the other subgroup 2 type of elements with frequencies from 19% up to ~ 50% and were reported for completeness but not considered for phylogenetic purposes.

The identified key substitutions were then investigated also in the processed pseudogenes HERV-W dataset, where their strong relation with the sequences distribution in the NJ trees was confirmed for the first 5 positions (96-100% frequency in subgroup 2 vs 0-3,5% in subgroup 1), while the last two mutations were shared among about the 75% of sequences (Table 7). Due to the pseudogenic LTRs truncated structure, the subgroup division was evident in the 3′LTRs tree (U3-R, positions 1-326 in LTR17) where 5 key positions out of 7 are maintained. The pseudogenic 5′LTRs (R-U5, positions 256-780) harbor instead only the two less represented key positions and showed a more confused topology, underlining the importance of the described substitutions in the phylogenetic asset of the group.

Table 7. *Recurrent mutations in HERV-W subgroup 2 LTRs*

Position [a]	**Substitution[b]**	**Frequency[c]**			
		PV* subgr. 2	PG° subgr. 2	Solitary subgr. 2	subgr. 1
43	C>T	100	100	98	0,7
95	C>T	100	95,8	96	3,4
100	T>C	97,3	100	95	2,2
180	C>T	97,3	100	95	0
254	A>G	97,3	96	87	1,4
706	A>G	97,4	73,3	88	1,7
765	G>A	95	73,3	90	1,7
Type 2A additional mutations					
Position	Substitution	PV* subgr. 2A	PV* subgr. 2B	Subgr. 1	
456	C>T	100	41,7	10,5	
498	A>G	92,6	33,3	3,5	
Type 2B additional mutations					
Position[a]	Substitution[b]	PV* subgr. 2B	PV* subgr. 2A	Subgr. 1	
133	A>G	100	48,1	0	
188	C>A	90	19,2	0	
252	C>G	90	40,7	1,7	

[a] Nucleotide positions are referred to RepBase Update LTR17 consensus

[b] Substitutions are indicated with the original and the acquired nucleotide separate by >

[c] Relative percentage based on the total sequences that hold the position in an alignment

*Proviruses °Processed Pseudogenes

3.4.3 Extended phylogenetic analysis of HERV-W genomic LTRs

Considering the relevance of LTR structural characteristics for HERV-W classification purposes, we retrieved via Genome Browser BLAT about 800 HERV-W LTRs present in hg19 assembly. This wider dataset has been used to assess the global reliability of the subgroup definition. The NJ tree analysis performed supported our classification, with a tree that resembled the topology observed for proviral and pseudogenic LTRs (Figure 10) and showed comparable distribution of solitary elements between the subgroups (71% classified as subgroup 1 and 29% as subgroup 2). When investigated for recurrent substitutions, the key positions defined for subgroup 2 were confirmed as commonly shared in 87-98% of the subgroup members and rarely present (1-6%) in the rest of the HERV-W LTRs dataset (Table 7).

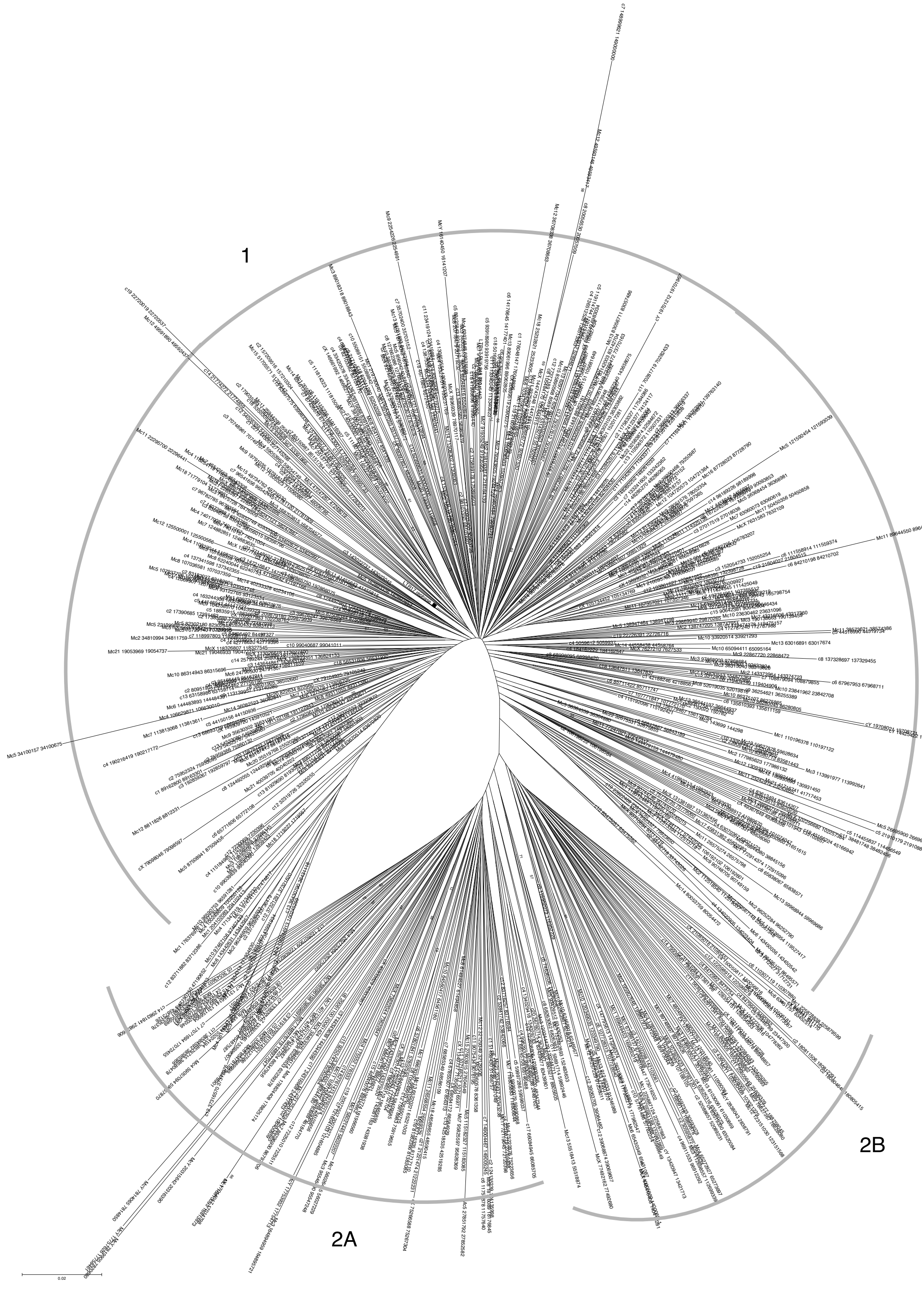


***Figure 10** Results from phylogenetic analysis of ~800 HERV-W LTRs sequences retrieved from hg19 assembly*
The evolutionary relationships were inferred by using the Neighbor Joining method based on the p-distance model and applying the pairwise deletion option. Phylogenies were tested by using the Bootstrap method with 500 replicates. Length of branches indicates the number of substitutions per site.

3.4.4 *gag, pol and env genes phylogenetic analysis*

The trees built for the retroviral *gag, pol* and *env* genes (Figure 11, panels A, B and C, respectively) did not showed the presence of any subgroup, and the nucleotide analysis confirmed that sequences share a comparable grade of homology. This result demonstrated that all the relevant HERV-W group phylogenetic variations are located in LTR sequence.

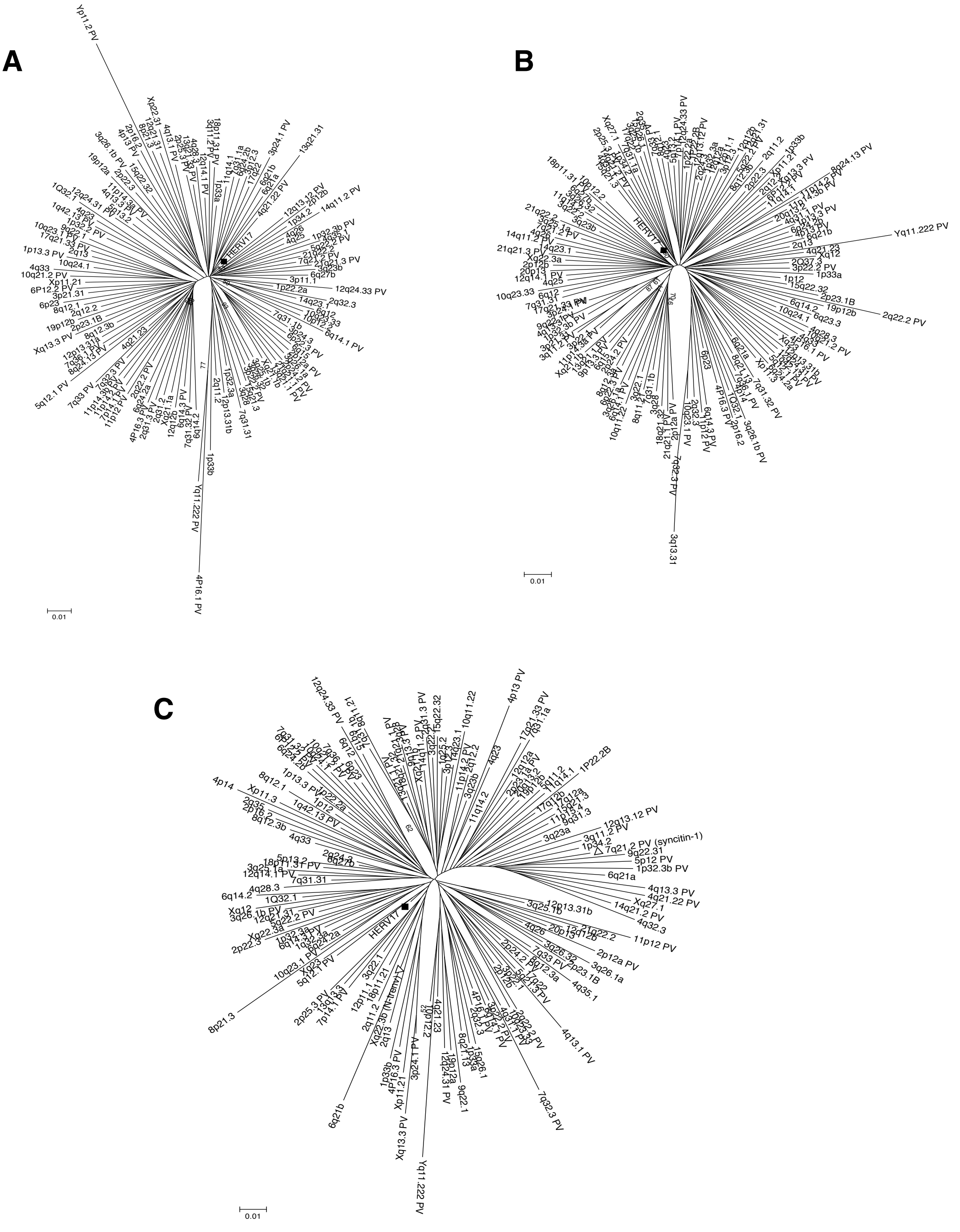
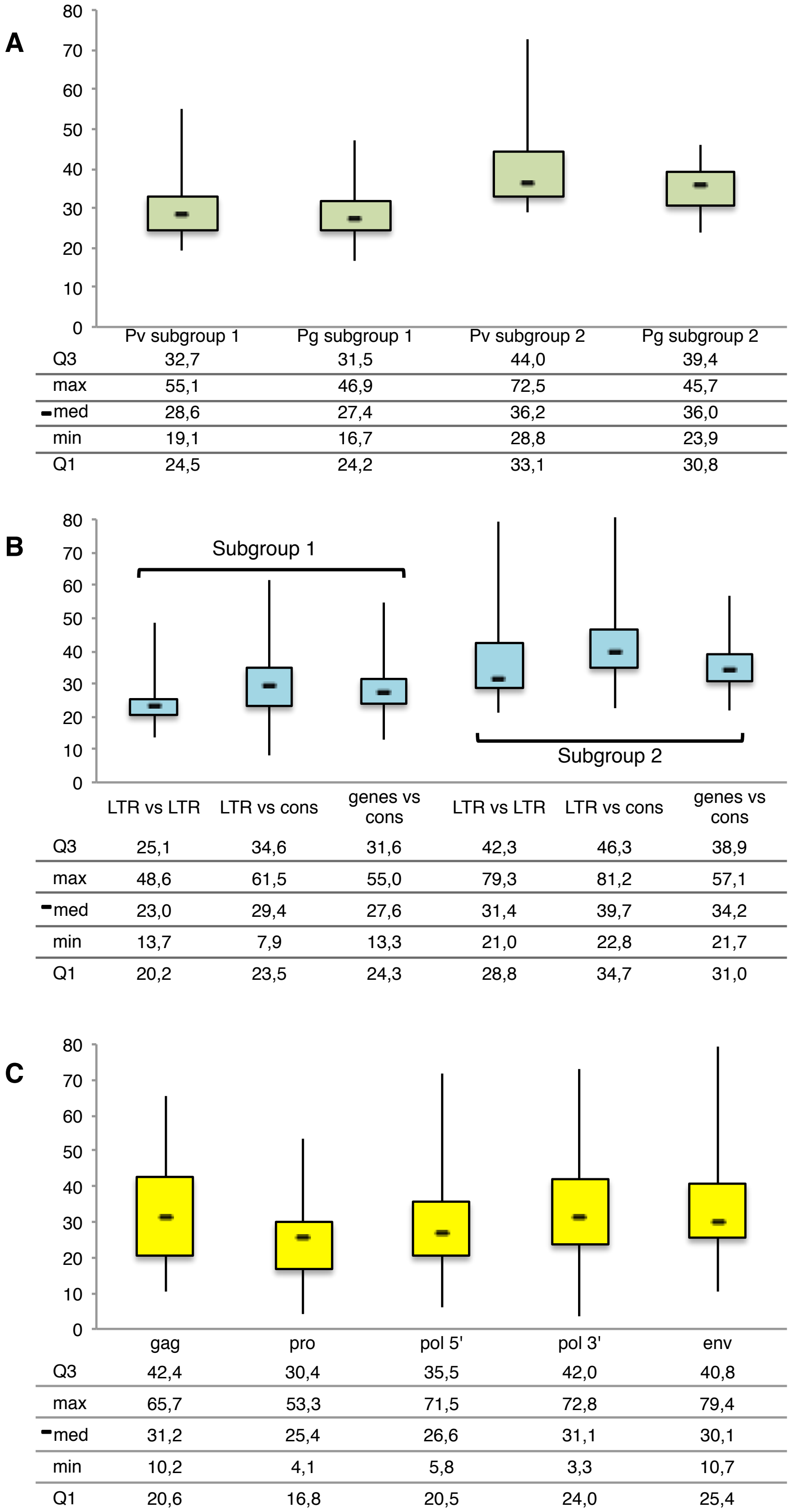


Figure 11. *Results from phylogenetic analysis of HERV-W genes nucleotide sequences.*
The evolutionary relationships of gag (A), pol (B) and env (C) genes were inferred by using the Neighbor Joining method based on the p-distance model and applying the pairwise deletion option. Phylogenies were tested by using the Bootstrap method with 500 replicates. Length of branches indicates the number of substitutions per site. HERV17 RepBase consensus is labeled with a black square

A LTR-based classification was previously suggested by Costas, that identified three distinct HERV-W subfamilies named 1, 2 and 3, on the basis of nucleotide differences described in a shorter version of the 3'LTR, with a truncation in correspondence to position 326 of LTR17, typical of pseudogenes [34]. Our data indicate instead that the HERV-W main subgroups are only two: subgroup 1 (associated to Costas subfamily 3) and subgroup 2 (related to Costas subfamilies 1 and 2). Subgroup 2 key mutations enclose the 5 mutations observed by Costas plus 2 more in the 3'LTR terminal portion. With respect to the previous classification, the one we propose is primarily based on a phylogenetic analysis, corroborated by the presence of high frequency key positions found in both 5' and 3' full-length LTRs and confirmed for the first time in a comprehensive HERV-W solitary LTR dataset.

3.5 Time of integration

It is known that, at time of integration, the 5' and 3' LTRs of the same provirus are identical [254] and accumulate random substitution in an independent way. Hence, to assess the HERV-W group estimated age we assumed for the human genome a substitution rate of 0.13%/nucleotides/million year [220] and used this rate to assess an action of divergence on different retroviral portions within each HERV-W sequence. Based on this assumption, we calculated the percentage of divergent nucleotides (D) i) between the 5' and 3'LTRs of the same provirus; ii) between each LTR (proviral and pseudogenic) and a generated consensus for each subgroup and iii) between a 150-300 nucleotides region of each HERV-W internal element *gag, pro, pol* RT, *pol* IN and *env* genes (proviral and pseudogenic) and a generated consensus. Regarding the two consensus-based approaches, in consideration that the substitution rate acts randomly on each sequence, the subgroup-generated consensus should ideally represent the ancestral situation. The obtained divergence values were used to calculate the age of the HERV-W sequences. For all three approaches the calculation is based on the relation T= D/0,13%, where T is the estimated time of integration (in million years) and 0,13% is the applied genomic substitution rate per million year. For the divergence between 5'- and 3'LTR of the same sequence, the obtained T value was divided by a factor of 2, considering that each LTR evolved and accumulated mutations independently. The estimated time of integration of each HERV-W sequence is reported in Table 6 and has been calculated as the average value resulted from the different approaches of divergence calculation (Figure 12).



	Pv subgroup 1	Pg subgroup 1	Pv subgroup 2	Pg subgroup 2
Q3	32,7	31,5	44,0	39,4
max	55,1	46,9	72,5	45,7
▬med	28,6	27,4	36,2	36,0
min	19,1	16,7	28,8	23,9
Q1	24,5	24,2	33,1	30,8

	LTR vs LTR	LTR vs cons	genes vs cons	LTR vs LTR	LTR vs cons	genes vs cons
Q3	25,1	34,6	31,6	42,3	46,3	38,9
max	48,6	61,5	55,0	79,3	81,2	57,1
▬med	23,0	29,4	27,6	31,4	39,7	34,2
min	13,7	7,9	13,3	21,0	22,8	21,7
Q1	20,2	23,5	24,3	28,8	34,7	31,0

	gag	pro	pol 5'	pol 3'	env
Q3	42,4	30,4	35,5	42,0	40,8
max	65,7	53,3	71,5	72,8	79,4
▬med	31,2	25,4	26,6	31,1	30,1
min	10,2	4,1	5,8	3,3	10,7
Q1	20,6	16,8	20,5	24,0	25,4

Figure 12. *Boxplot representations of HERV-W subgroups divergence based estimated period of integration.* The approximated age (in million years) was calculated considering the divergence values between the 5′ and 3′LTRs of the same provirus; between each proviruses and processed pseudogenes LTR and a generated consensus for each subgroup and between a 150-300 nucleotides region of each HERV-W proviruses and processed pseudogenes *gag*, *pro*, *pol* RT, *pol* IN and *env* genes and a generated consensus. Panel A: averaged values of age obtained with the three methods for each subgroup proviruses and processed pseudogenes; Panel B: single method estimations for the two HERV-W subgroups; Panel C: highlight of the heterogeneous action of the divergence at different genic regions.

In particular, the estimated time of integration of proviruses and processed pseudogenes sequences for both subgroups 1 and 2 (Figure 12, panel A) describes for the first time the HERV-W dynamic of insertion into the human genome, suggesting that: i) the first HERV-W integrations involved subgroup 2 and occurred more than 40 million years ago, with a diffusion of proviral and pseudogenic sequences until about 30 million years ago; ii) HERV-W subgroup 1 sequences are significantly younger with respect to subgroup 2 members (p<0,0005), and have been acquired mostly between 35 and 25 million years ago, occurring in average about 8 million years later than subgroup 2; iii) it is interesting to note that, for both subgroups, the dissemination of proviruses and processed pseudogenes took place virtually simultaneously. Moreover, despite both subgroups proviruses were processed by the LINE machinery to generate pseudogenes, the mechanism was more frequent for subgroup 1 proviruses (1:2,5 ratio with the number of related pseudogenes) than for subgroup 2 integrated elements (1:1 ratio). The reason for this is at the moment unclear. We attempted to connect the single processed pseudogenes to the original generating proviruses by a phylogenetic analysis of LTRs and major genes, expecting that the pseudogene elements could cluster with their respective HERV-W locus of origin. However, the great majority of pseudogenes clustered with different proviral loci according to the sequence portion considered (data not shown). This result, together with the estimated time of diffusion of processed pseudogenes, suggests that these elements acquired a comparable amount of heterogeneity since their mobilization by L1 elements, and it is thus not possible to univocally assign them to a single provirus.

It is important to note that the traditional sole comparison of the two LTRs of the same proviral sequence would not be sufficient for a reliable estimation of the group dynamics of diffusion within the human genome. In fact: i) the LTR vs LTR method could not be applied at all to processed pseudogenes, due to the short region in common between the 5′ and 3′LTRs, and undefined elements, allowing thus to estimate the time of integration for proviruses only (36% of the total HERV-W sequences); ii) also in the case of proviral sequences, the lack of one or both LTRs make effectively possible such calculation only for the 70% of proviruses (23% of the total HERV-W members). The two additional approaches completed and improved the time of integration estimation, allowing to consider a larger subset of elements (94% of the total HERV-W members) and to represent also the truncated processed pseudogenes and older and less intact sequences, which were not previously taken into account. Importantly, the combination of multiple divergence calculations provided significant improvements also in age estimation reliability and precision. The expression of each HERV-W sequence time of integration through the use of an averaged value allowed to determine the standard deviation and to reduce estimation biases related to outliers and different selective pressure that are reported to interest LTR elements with respect to the rest of the retroviral genome [239] (Figure 12, panel B). Data showed that some proviruses had a 0,3 - 2 folds higher age estimation when calculated using the LTR vs consensus method as compared with the LTR vs LTR method. Despite the absence of a clear explanation, it is possible to speculate that the exogenous viruses that gave rise to these sequences harbored some nucleotide differences in their LTRs that are not properly represented in the consensus sequence, built on the majority of viruses, leading to an apparent higher amount of mutations. In addition, data showed a higher divergence in the *gag*, *pol*-3′ (including IN) and *env* portions, leading to a older age estimation with respect to the internal *pro* and *pol*-5′ regions (Figure 12, panel C), thus suggesting different mutation rates according to the specific viral portions.

Taken together, these results suggest that the HERV-W group integration started ~ 40 million years ago at the time of the *Catarrhini* primates, after the divergence between *Platyrrhini* and Old World Monkeys. This is in line with previous studies [34, 255, 256], which were based either on the presence, in different Old World Monkeys samples, of HERV-W *pol* PCR products [255] or on Southern Blot DNA analysis [256], using in both cases MSRV-derived primers and probes, respectively; or on the divergence among HERV-W subfamilies [34]. Overall, these studies gave thus just a general rough overview of HERV-W group acquisition by primates genomes, without information about the single members diffusion and the precise time period of activity of the group. In the present study, the time of insertion has been estimated for each single HERV-W locus through at least two different methods of age calculation, providing a precise and exhaustive picture of the group diffusion among primates, with a rather long period of activity that took from place from ~ 40 million years ago until approximately 20 million years ago.

The estimated age of the single HERV-W sequences was generally also supported by the identification of each HERV-W orthologous locus in primates until the Oldest Common Ancestor (O.C.A.) (Table 6). Results showed that the great majority of sequences are shared from human to Rhesus Macaque (61%) or to Gibbon (31%), with an entry that must be occurred after their separation from the *Platyrrhini* parvorder (43 million years ago) and before their divergence from the evolutionarily younger hominoids, occurred around 30

(Rhesus Macaque) and 20 (Gibbon) million years ago [221]. Few elements were also found starting only since Orangutan (12), Gorilla (3) and Chimpanzee (2) (Table 6), but in these cases the estimated age was higher than expected. This probably suggests that such sequences were lost in older primates, even though their absence in Rhesus and Gibbon could be also due to a lower efficiency of Genome Browser comparison between the human genome and the most ancient *Catarrhini* assemblies. Finally, a single HERV-W element was found only in the human genome, on locus 12q13.3. This data is unexpected because no human specific HERV-W elements have been reported so far, but could not be supported by reliable age estimation due to the shortness of the sequence (about 1500 nucleotides) and the lack of both LTRs.

3.6 Structural features

As previously described in detail, beside the traditional PBS type and the by now widely used *pol* similarities for taxonomic and classification purposes, some other structural traits can be also highly useful to better understand retroviral phylogeny [257]. Such characteristics features, in the case of Gammaretroviral ERVs, includes i) the number of Gag NC zinc finger motifs, ii) the presence of a GPY/F motif in the Pol IN C-terminal domain, and iii) the nucleotide compositional bias.

3.6.1 PBS type

The PBS type has been historically used to identify the different HERV groups that were commonly designated with the amino acid single letter of the corresponding tRNA. Currently this nomenclature is not considered a sufficient and reliable taxonomic marker, especially because it is not based on HERV phylogeny [6, 8]. In the analyzed HERV-W elements, the PBS was present in 111 sequences and was located approximately 4 nucleotides downstream the 5′LTR (from nucleotide 4 to 21 in HERV17 consensus) (Figure 5). The PBS type of the single HERV-W sequences is reported in Table 6, while a graphical overview of the PBS types found in the entire HERV-W dataset and in each subgroup is provided in Figure 13. In general, Tryptophan (W) was, as expected, the most common PBS type: it was found in a total of 60 sequences, representing about the 58% of the identified HERV-W PBSes. Therefore, noteworthy, about half of HERV-W elements analyzed had a non-W PBS type, confirming the relatively low taxonomic value of this feature. Particularly, Arginine PBS was rather common (R, 21), followed by Phenylalanine (F, 9), Isoleucine (I, 4), Serine (S, 3) and Proline (P, 2). Other PBS types found in single HERV-W sequences were moreover Leucine (L), Asparagine (N), Glutamic Acid (E) and Glycine (G). In the remaining 8 elements, the PBS sequence was present but it was not possible to unambiguously classify it.

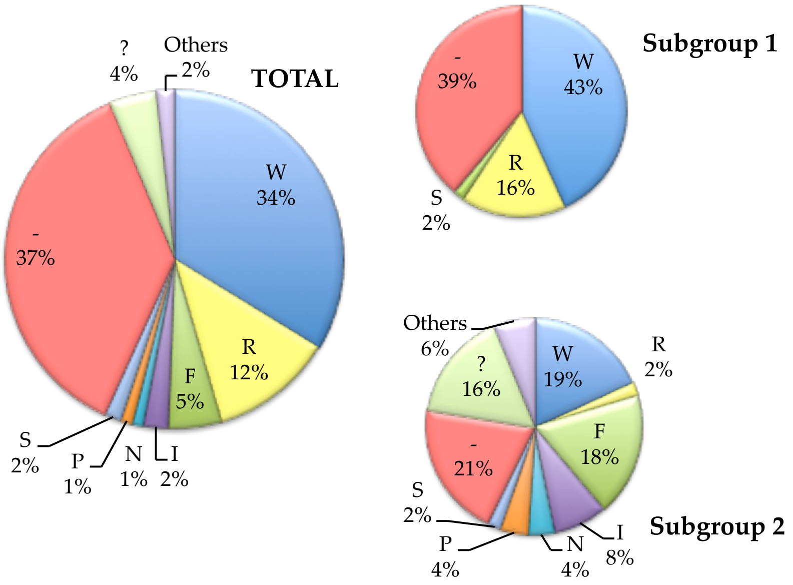


Figure 13. *PBS types among all HERV-W sequences and diversity between subgroup 1 and subgroup 2.* The PBS types are identified by the amino acid single letter of the corresponding cellular tRNA. W= tryptophan, R= arginine, F= phenylalanine, I= isoleucine, S= serine, P= proline. "Others" category encloses Leucine (L), Glutamic Acid (E) and Glycine (G), each found only in one sequence. Elements that lost the PBS sequence (-) or with PBS with ambiguous assignation (?) are also included.

Regarding the various PBS frequencies in the two HERV-W subgroups, interestingly, subgroup 1 elements retaining the PBS sequence showed a more homogeneous situation, presenting almost the 100% of W or R as putative tRNA used. This was expected, since the W codon is the most commonly associated to the HERV-W group while the R one differs only slightly from it, and sometimes the two codons may overlap due to a single nucleotide shift in the PBS sequence [8]. Differently, subgroup 2 elements revealed a more heterogeneous PBS type usage, including all the unusual tRNA sequences and all the ambiguous PBSes with no clear assignment. These atypical PBSes are probably derived from the accumulation of several substitutions, in accordance with the older appearance and the longer permanence of these sequences in primates genome. To summarize the general variation of the PBS sequence within HERV-W group we generated a logo (Figure 14, panel A) in which the letter height is proportional to the nucleotides conservation at each position. As expected, the TGG starting nucleotides, which are shared by almost all the PBS types, were the most conserved among the 18 bases analyzed. Interestingly the middle portion of the sequence showed a high variability, especially at positions 4-6 and 11, indicating a rather large diversity of PBSes in HERV-W group.

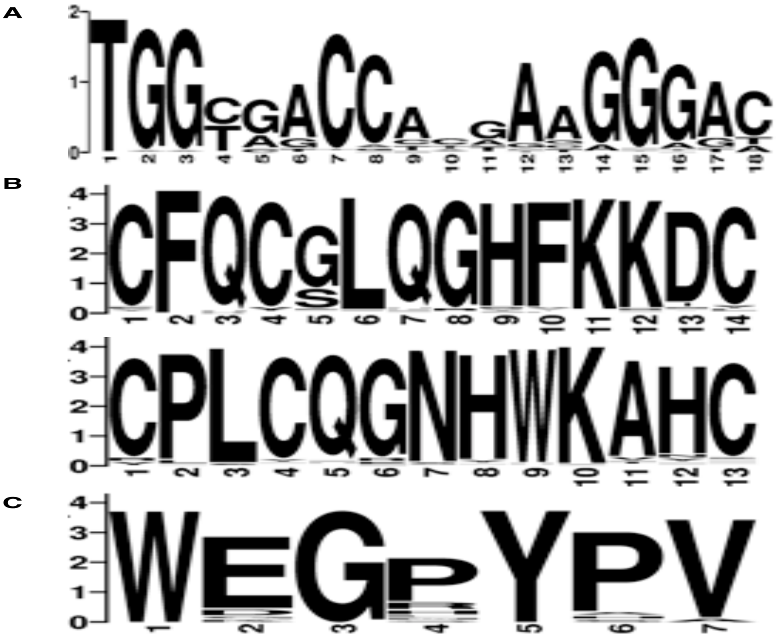


Figure 14. Logos representing the HERV-W main structural features conservation.
Overall height indicates the sequence conservation, while the letter height indicates the relative frequency of each nucleotide/amino acid. Created at http://weblogo.berkeley.edu/logo.cgi
Panel A: PBS nucleotide sequence; Panel B: Gag NC Zinc fingers amino acid motifs; Panel C: Pol IN GPY/F amino acid motif.

3.6.2 Gammaretroviral features

Beside the PBS type, we identified and analyzed also those structural features typically shared among retroviral sequences within the same genus, that can be used as taxonomic and phylogenetic markers [257]. As previously reported [257], the main Gammaretroviral features are i) one Gag NC Zinc finger motif, involved in the retroviral RNA interaction during packaging [258]; ii) the C-terminal Pol IN GPY/F motif, that binds the host DNA and could have a role in the integration specificity [259, 260] and iii) a nucleotide frequency bias, determined by the action of encapsidated host RNA editing systems [257].

The Gag NC Zinc finger, corresponding to nucleotides 4021-4062 in the RepBase assembled LTR17-HERV17-LTR17 reference sequence (Figure 5), has a typical $CX_2CX_4HX_4C$ amino acid motif. It was found in almost all sequences that retained the harboring genetic region, with a higher prevalence in proviral sequences, being however the most complete in term of genetic composition. Moreover, noteworthy, a second Gag NC Zinc finger was identified in 96% of the sequences that retained the harboring genetic region (nucleotides 4093-4130) (Figure 5). This second Zinc finger has a modified structure with respect to the usual one, showing the loss of one of the variable residues ($CX_2CX_3HX_4C$). The amino acid composition of both motifs was highly conserved (Figure 14, panel B). The presence of a second Zinc finger was not previously reported for HERV-W group, and its structure is consistent with the second Zinc finger found in a subset of HERV-H sequences, another Gammaretroviral group [261]. However, while for HERV-H a correlation between the presence of this second motif and the age of sequences was proposed, for HERV-W we could not observe such correlation (data not shown).

The Pol IN domain contains a GPY/F motif, a stretch of conserved amino acids with the general WX_nGPYXV structure corresponding to nucleotides 7501-7521 in the reference sequence (Figure 5). Considering that the C-terminal portion of the *pol* gene was deleted in 85% of sequences, in the remaining few members the GPY/F feature was found with a 100% frequency. Also this feature showed a conserved amino acid sequence (Figure 14, panel C)

Regarding the nucleotide composition, HERV-W sequences present an overall weak bias in purines, tending to be richer in Adenine (A, ~ 30%) and poorer in Guanine (G, ~ 22%) (data not shown). Among Gammaretroviruses, an impoverishment of G nucleotide was previously observed for HERV-H group in association with an higher content of Cytosine [261], while the G to A hypermutation condition was reported for HERV-K group [262] and is a well known effect of the APOBEC3 defensive action against HIV-1 Lentivirus [263]. Hence, it is possible to speculate that this editing system could have played a role as a control mechanism to limit HERV-W and other endogenous elements mobility during evolution [264], also considering APOBEC3 ability to greatly inhibit the LINE mediated transposition of other retroelements [265].

3.7 Genomic context of insertion

The current major field of HERVs investigation is surely the different groups general expression and coding capacity. However, the impact of these sequences on the host greatly depends also on their genetic surrounding. The context of integration can, in fact, modulate HERVs physiology, and HERV sequences inserted in proximity of human genes are known to be able to influence their expression [240–242, 244, 246, 249, 250], even in the absence of an expressed product. As reported for other HERV groups [266], the analysis of the genomic context of all 213 HERV-W confirmed that the majority of sequences are located in intergenic regions, with the exception of 80 elements inserted into human coding and non-coding genes, that were further characterized.

3.7.1 HERV-W sequences inserted into human coding genes

55 elements (69% of HERV-W sequences into human genes, 26% of the total sequences) are inserted into human coding genes, mostly exclusively into intronic regions (53/55) (Table 8). These elements showed a strong anti-sense orientation with respect to the surrounding gene (43/55) that was more frequent for pseudogenic sequences than for proviruses (84% and 67%, respectively). The fact that HERV-W intronic elements are present mostly in antisense orientation could reflect a bias due to an evolutionary and post-insertional selection [267]. Noteworthy, while most of the identified sequences were already characterized for their genomic context [119, 268], 8 of them are reported for the first time to be inserted in human genes. Based on Genome Browser annotations, pseudogene 21q22.2 resulted from an overlap in antisense orientation with the first two exons of IGSF5 gene, that produces a protein involved in junction and adhesion formation, and with a corresponding IGSF5 highly similar mRNA found in placental tissues (AK092516). Interestingly, more than half of the 55 genes that held HERV-W sequences were reported to be positively associated with a disease or a pathologic trait, that in most cases affect the neurological and the cardiovascular systems (Table 8). Considering that the genomic context of integration and the orientation of HERV-W sequences appeared to influence their expression [268], the present mapping of those elements may aid in the understanding the potential effects of these integrations on human health and to direct further investigations to the genes involved.

Table 8. *HERV-W genomic context: insertions into human coding genes*

HERV-W	Human Gene	Gene/protein function and associations
1p34.2 (+)	*HIVEP3* Int 1(-)	Transcription factor, binds Ig and T-cell receptors recombination signal
1q25.2 (-)	*RASAL2* Int 1 (+)[2]	RAS superfamily of small GTPases protein activator like. Associations: BMI, weight
1q42.13* (-)	*ZNF678* Int 2 (+)[1] [2]	Zinc Finger protein. Associations: body height
2p23.1a* (+)	*LCLAT1* Int 2 (+)[2]	Predominantly remodels anionic phospholipids in Endoplasmic Reticulum
2p16.2 (+)	*ASB3* Int 1/2 (-)[2]	Suppressor of cytokine signaling proteins and their binding partners

2q22.2* (-)	*KYNU* Int 2 (+)[1] [2]	NAD cofactors biosynthesis from tryptophan. Associations: body height, cholesterol, schizophrenia
2q24.3 (-)	*COBLL1* Int 2 (-)[1]	Cordon bleu WH2 repeat protein-like 1. Associations: BMI, Cholesterol, HDL, triglycerides, stroke, response to statin therapy, anthropometric sexual dimorphism
2q31.2a (-)	***AGPS*** Int 1 (+)[2]	[603051] Mutations are cause of rhizomelic chondrodysplasia punctata type 3
2q35 (-)	*DIRC3* Int 1 (-)[2]	Disrupted In Renal Carcinoma long non-coding RNA. Associations: diabetes mellitus
3p22.2* (-)	*SLC22a14* Int 1 (+)[2]	Solute Carrier transmembrane protein
3q22.1 (-)	*NEK11* Int 14/13 (+)[2]	Never In Mitosis kinase. Involved in DNA replication and G_2/M checkpoint response to DNA damage. Related to embryonic lethality and preeclampsia
3q23b (+)	*XRN1* Int 1 (-)[2]	Exoribonuclease involved in Long noncoding RNA decapping and miRNA regulation
3q26.32 (+)	*ZMAT3* Int 2/3 (-)[1] [2]	Zinc Finger Matrin. Acts as a bona fide target gene of p53/TP53
4p16.3* (-)	*ZNF595* Int 3 (+)	Zinc Finger Protein. Function as transcription factor
4p16.1* (+)	*ACOX3* Int 1 (-)[1] [2]	Oxidizes the CoA-esters of 2-methyl-branched fatty acids
4q31.3 (+)	***ARFIP1*** Int 2 (+)[2]	ADP ribosylation factor interacting protein1. [605928] Enhance the cholera toxin activity
5q12.1* (+)	*DEPDC1B* Int 2 (-)[2]	Significantly upregulated in nonsmall cell lung carcinoma cell lines (reduced patient survival)
5q22.2* (+)	*ACOT13* Int 1 (+)	Acyl-CoA thioesterase. Involved in regulation of lipid composition and metabolism
6q12 (+)	***EYS*** Int 13 (-)[2]	[612424] In photoreceptor layer: mutated in autosomal recessive retinitis pigmentosa
6q14.3a* (-)	*TBX18* Int 7 (-)[1] [2]	Role in embryonic development. Associations: cholesterol, coronary disease
6q21a (+)	*ATG5* Int 6 (-)[1] [2]	Autophagy related apoptosis specific protein. Associations: lipoproteins, LES
6q21b° (+)	***PDSS2*** Int 2 (-)[2]	Prenyl (decaprenyl) Diphosphate Synthase, Subunit 2. Synthesizes the side-chain of coenzyme Q. [610564] Coenzyme Q10 deficiency, primary, 3: fatal encephalomyopathy and nephrotic syndrome
6q21c (+)	*SLC16A10* Int 1 (+)[1] [2]	$\text{Na}^{(+)}$-independent transport of aromatic amino acids across plasma membrane. Associations: cholesterol, LDL
6q24.2a (-)	*AIG1* Int 1 (+)[2]	Androgen-induced. Associations: C-reactive protein, insulin, myocardial infarction
7p21.1 (-)	*BZW2* Int 3 (+)	Homo sapiens basic leucine zipper and W2 domains 2
7p14.1* (-)	***SUGCT/C7orf10*** Int 1 (+)[2]	[609187] Mutations are associated with glutaric aciduria type III. Others: BMI, fat distribution, cardiomegaly, coronary disease, pancreatic and prostatic neoplasms
7q31.1a (+)	*NRCAM* Int 2 (-)[1]	Neuronal Cell Adhesion Molecule. Associations: autism, obsessive compulsive disorder, schizophrenia
7q31.1b (-)	***FOXP2*** Int 2 (+)[1] [2]	[605317] Required for development of speech and language regions of the brain during embryogenesis. Associated to speech-language disorders
8p21.3 (+)	*SLC18A1* Int 10/11 (-)[2]	Involved in vesicular transport of biogenic amines. Associations: bipolar disorder, major depressive disorder
8q12.3a (-)	*NKAIN3* Int 3 (+)	Na+/K+ transporting ATPase Interacting proteins. Associations:

		mental competency, neuroblastoma, stroke
8q12.3b (+)	***CYP7B1*** Int 1 (-)[2]	[603711] Cyp450 enzyme. Associations: bile acid synthesis congenital defect, spastic paraplegia. Others: Alzheimer disease, lipoproteins, schizophrenia
8q21.11 (+)	***UBE2W*** Int 2(-)[2]	Ubiquitin-conjugating enzyme. Along with ubiquitin-activating (E1) and ligating (E3) enzymes, coordinates the ubiquitin addition to proteins. [614277] Interacts with FANCL and regulates the monoubiquitination of Fanconi anemia protein FANCD2
8q21.13 (+)	*ZNF704* Int 2 (-)[2]	Zinc Finger protein
9p24.1 (+)	***PTPRD*** Int 12 (-)[2]	Protein tyrosine phosphatase, receptor type, D. [601598] Restless Legs Syndrome. Associations: asthma, BMI, cholesterol, lipids, lipoproteins, triglycerides, diabetes.
9p13.3* (-)	*CD72* Int 1 (-)[1] [2]	B-cell proliferation and differentiation antigen. Associations: lupus erythematosus
10q23.33 (-)	***CYP2C19*** Int 6 (+)[2]	[124020] Cyp450 enzyme, responsible for therapeutic agents metabolism. Associated to metabolic defects and variants
10q24.1 (-)	***ENTPD1*** Int 1 (+)[2]	[601752] Triphosphate Diphosphohydrolase. Associated with Spastic Paraplegia
11p14.2* (-)	***ANO3*** Int 14 (+)[1] [2]	[610110] May act as a chloride channel. Associations: Dystonia 24. Others: bmi, obesity, c-reactive protein, cholesterol, coronary disease, schizophrenia
11q14.1 (-)	*AAMDC* Int 2 (+)[2]	Adipogenesis Associated Mth938 Domain Containing
11q14.2 (-)	*PRSS23* Int 2 (+)[2]	Encodes a conserved member of the trypsin family of serine proteases
12p13.31b (-)	*RIMKLB* Int 5 (+)[2]	Catalyses ATP-dependent condensation of NAA and glutamate to produce NAAG
12q23.3 (+)	*SLC41A2* Int 1 (-)	Solute carrier family 41member 2
13q13.3 (+)	*ALG5* Int 7/8 (-)[2]	Participates in N-linked glycosylation of proteins
14q11.2* (+)	*TCRA* Int 1 (+)[2]	T cell receptor alpha locus
14q21.2* (-)	*FAM179B* Int 7 (+)	Homo sapiens family with sequence similarity 179 member B
14q23.1 (+)	*C14orf37* Int 4 (-)[2]	Associations: attention deficit disorder with hyperactivity
17q12a (+)	*SLFN14* Int 3 (-)[2]	Implicated in regulation of cell growth and T-cell development (studies in mouse
17q12b° (-)	*ACACA* Int 2/6 *(-)*[2]	Biogenesis of long-chain fatty acid. Associations: BMI, breast cancer
17q22 (-)	*STXBP4* Int 8 *(+)*[2]	Translocation of transport vesicles from cytoplasm to plasma membrane, like the insulin-stimulated GLUT4 translocation in adipocytes. Associations: BMI, cholesterol
19p12a (+)	***ZNF90*** Int 1 (+)[2]	Zinc finger protein 90. May be involved in transcriptional regulation. [603973]
19q13.2a (+)	*ZNF780A* Ex 9 (-)[2]	Zinc finger protein 780A
19q13.2b (-)	*CYP2A7* in 1 (-)[2]	Cytochrome P450, family 2, subfamily A, polypeptide 7
21q22.2 (-)	*IGSF5* Ex 1-2, Int 1 *(+)*[2]	Participates at tight-junctions (kidney, gut) or acts as adhesion molecule (testis). Associations: coronary disease, lipoproteins, Parkinson disease, stroke
Xp11.21 (-)	*FAAH2* Int 7 (+)[2]	Degradation and inactivation of bioactive fatty acid amides
Yq11.222* (+)	*CD24* Int 1 (-)	Mature granulocytes and B cells surface antigen

Proviruses and undefined sequences are labeled respectively with * and °. For HERV-W sequences and genes, the strand direction is reported into round brackets. Bold genes are listed as OMIM diseases associated and the relative accession number is reported into square brackets. Underlined genes are reported to be positive associated with specific phenotypes in UCSC Gene annotations.

[1] Already reported in Li et al. 2011. [2] Already reported in Schmitt et al. 2013

3.7.2 HERV-W sequences inserted into human non-coding genes

In addition, 25 HERV-W loci (31% of HERV-W sequences into human genes, 12% of the total sequences) were integrated into 29 human non-coding genes, of which the great majority (22) is reported here for the first time (Table 9).

Table 9. *HERV-W genomic context: insertions into human non-coding genes*

HERV-W	**Human Gene**	**Gene function and associations**
1p12 (-)	*LOC101929147* Int 4 (+)	Uncharacterized antisense long non-coding RNA
1p13.3* (-)	*TCONS_00000271* Int 3 (+)	Large intergenic non coding RNA
1q32.1 (-)	*LOC284581* Int 1 (+)[2]	Uncharacterized antisense long non-coding RNA
2q11.2 (-)	*STARD7-AS1* Int1 (+)[2]	StAR-related lipid transfer domain protein 7 antisense long non coding RNA (LOC285033**)**
2q24.3 (-)	*TCONS_00004484* Int 1 (-)	Long intergenic non coding RNA
2q31.2b (+)	*MIR548N* Int 1 (+)[2]	Homo sapiens microRNA 548n
3q25.1b (+)	*CLRN1-AS1* Int 1 (+)	CLRN1 antisense non-coding RNA
4p13* (-)	*TCONS_00007753* Int 1 (-)	Long intergenic non coding RNA
4q23 (-)	*LOC100507053* Int 1 (+)	Uncharacterized antisense long non-coding RNA
4q28.3 (+)	*TCONS_00007833* Int 1 (-)	Long intergenic non coding RNA
4q32.3 (+)	*MIR5684* Int 2 (+)	MicroRNA (post-transcriptional regulation of gene expression)
6q15 (-)	*TCONS_00011526* Ex 1, Int 1 (-)	Long intergenic non coding RNA
6q27a° (+)	*TCONS_l2_00024517* Int 2, Ex 3 (+) *TCONS_l2_00024518* Int 1, Ex 2 (+) *TCONS_l2_00024519* Int 1 (+)	Long intergenic non coding RNAs
7p14.2* (+)	*DQ594967* Ex 1(-)[2]	Antisense non coding RNA
8q12.1 (-)	*TCONS_00015019* Int 1 (-) *AC022555.1* (-)	Long intergenic non coding RNA Pseudogene
9p21.3 (+)	*LOC441389* Int 5 Ex 6 (+)[2]	Uncharacterized long non-coding RNA
10q11.22 (-)	*TCONS_00017977* Int 1 (-)	Long intergenic non coding RNA
11q14.2 (-)	*PRSS23* Int 2 (+)	Protease serine 23 near-coding RNA
11q23.3 (-)	*TMPRSS4-AS1* Int 2 (-)[2]	Antisense non-coding RNA
13q21.33* (+)	*LINC00383* Ex 1, Int 1 (+)	Long intergenic non coding RNA
13q31.3° (+)	*TCONS_00021873* Int 2 (+)	Long intergenic non coding RNA
21q21.1* (-)	*MIR548XHG* Ex 1, Int 1 (-)	MIRNA548X host gene long non-coding RNA
14q22.1 (+)	*AL163953.3* Int 3 (+)	Long non-coding RNA
19p12d (+)	*AK125686* Int 2 (-)[2]	Antisense non coding RNA
Xq13.3* (-)	*TCONS_00016997* Ex 1-2, Int 1 (+) *AL451105.1* (-)	Long intergenic non coding RNA Pseudogene

Proviruses and undefined sequences are labeled respectively with * and °. For HERV-W sequences and genes, the strand direction is reported into round brackets.

[1] Already reported in Li et al. 2011. [2] Already reported in Schmitt et al. 2013

These elements were mostly inserted into regions associated with the production of long non-coding RNA (lincRNA) and microRNA (miRNA), regulatory molecules that operate on different levels of gene expression. These HERV-W proviruses (6), pseudogenes (17) and undefined sequences (2) showed different characteristics with respect to the HERV-W loci located into coding regions. Firstly, despite even in this case a majority of intron localization (26) was observed, 8 HERV-W sequences were co-localized with 9 exons, frequently situated at the transcriptional start site of the non-coding gene. Secondly, in this case the antisense bias was not present, since 19 out of the 29 non-coding genes showed the same orientation with respect to the HERV-W inserted elements. These observations suggest that the LTRs of these HERV-W could provide regulatory signals for lincRNA, as already highlighted for HERV LTRs in general [269]. The great majority of non-coding genes harboring HERV-W sequences were still uncharacterized, but some elements were reported as being associated with post-transcriptional regulation (MIR5684) or related to proteins involved in lipid transfer and proteolytic activity (STARD7-AS1 and PRSS23). Overall, the percentage of HERV-W sequences inserted into human genes (38%) is higher than the percentage of bases spanned by human genes (24%) [270, 271] suggesting that integration events could have been biased for genic or against intergenic regions.

3.7.3 HERV-W sequences predicted binding to cellular transcription factors

Finally, the genetic context of HERV-W elements was evaluated for the possibility to bind cellular TFs that normally interact with the host DNA. The analysis was based on the data obtained through ENCODE Transcription Factors ChIP-seq and Factorbook databases, and only TFs with higher score (from 800 up to 1000) were considered (Table 10). Results showed that 16 HERV-W elements could tentatively bind cellular TFs, including POLR2A, MAFK, E2F1 and TCF7L2. 12 of these elements were proviruses, and 7 of them plus 1 pseudogene were inserted into human coding genes. The higher representation of proviral sequences is probably due to the presence of complete LTR structures, retaining sites deputed to TF recognition. Despite the fact that the detection of predicted TF binding sites is not sufficient to suggest a possible transcription, their presence in HERV-W elements that are co-localized with human genes could potentially have an effect on the transcription of such genes and need to be further investigated at the expression level.

Table 10. *HERV-W genomic context: transcription factor (TF) binding sites*

HERV-W	TF recognized	Position	Score (0-1000)
2p12a*	POLR2A	chr2:76098843-76099352	803
2q22.2*	E2F1	chr2:143661226-143661546	958
3p22.2*	CTCF	chr3:38331061-38331485	900
4p16.1*	POLR2A	chr4:8429472-8430544	1000
	TCF7L2	chr4:8424096-8424592	922
6p12.2*	TFAP2C	chr6:52783052-52783485	1000
	FOXA2	chr6:52783244-52783462	848
	STAT3	chr6:52784270-52784579	808
6q14.3*	STAT3	chr6:85427859-85428174	1000
	CEBPB	chr6:85427862-85428118	817
7q21.1*	TCF7L2	chr7:92103429-92103733	1000
7q31.1a	TCF7L2	chr7:107981247-107981830	1000
	E2F1	chr7:107981308-107981897	1000
7q33*	YY1	chr7: 34270591-134271127	1000
7q36.1*	FOXA1	chr7:149370177-149370408	1000
9q22.1	TCF7L2	chr9:91556701-91556965	1000
10q21.2*	GATA3	chr10:62797340-62797529	1000
	E2F1	chr10:62796837-62797697	806
10q23.1*	GATA1	chr10:86284672-86285183	1000
	MAFK	chr10:86285572-86285911	1000
	MAFF	chr10:86285647-86285793	1000
	TBL1XR1	chr10:86284785-86285185	824
10q24.1	TAL1	chr2:97480654-97480806	928
	TEAD4	chr10:97480630-97480810	859
10q21.3	MAFK	chr10:65805045-65805364	1000
21q21.1*	TFAP2C	chr21:20128637-20128925	1000
	YY1	chr21:20131977-20132464	859

Data obtained from Genome Browser Encode Transcription Factor ChIP-seq database.
Proviruses and undefined sequences are labeled respectively with * and °.
Bold loci are inserted into human coding or non-coding genes.

3.8 Env putative proteins analysis

Due to its well established physiological role, the ERVWE1/Syncytin-1 ORF has been characterized in detail in terms of structure and functional domains, as recently reviewed [31]. Hence, we wanted to compare the Syncytin-1 precursor ORF and its features with respect to the most complete *env* genes found in our dataset, in order to predict the conservation of those sites reported to be involved in Syncytin-1 protein *in vivo* functions. Within our HERV-W elements, in addition to the Syncytin-1 locus ORF (7q21.2, ~ 1,6 Kb, 538 aa), we identified 16 full-length or near full-length *env* genes (> 1,4 Kb), 3 more than previously reported with similar parameters [272], and 10 conserved but shorter *env* genes (from 1398 to 801 nucleotides). The bioinformatics translation of these *env* genes led to obtain the correspondent putative proteins (puteins), showing a length range of 483-559 aa and 267-

466 aa, respectively (Table 11). These *Env* puteins were obtained from different reading frames in the often-damaged *env* gene candidates, and are thus just a bioinformatics model useful to evaluate the predicted domains structure.

Table 11. *Env puteins analysis*

483-559 amino acids (aa) Env puteins				*267-466 amino acids (aa) Env puteins*			
Sequence	**Length**	**Stops**	**Shifts**	**Sequence**	**Length**	**Stops**	**Shifts**
7q21.2*	**538**	**0**	**0**	11p15.4	475	2	3
1p32.3b*	559	3	2	3p24.1*	466	2	0
6q21a	552	4	2	9q31.3	462	3	1
15q21.3	543	2	2	3q23a	453	2	1
4q13.3*	542	0	2	*4q32.3*	443	1	2
5q11.2	542	0	4	1q32.3a	361	0	3
5q21.3*	542	6	2	5p12*	355	6	0
12q13.12*	542	2	1	1p34.2	352	2	1
14q21.2*	542	3	1	Xq27.1	352	4	1
Xq22.3b	542	1	0	4q21.22*	320	0	1
3q11.2*	541	1	3	17q12a	296	0	2
4q31.1*	541	2	2	*9q22.31*	267	1	0
17q12b	540	2	2				
Xp22.31	529	0	6				
20q13.2	483	0	1				

Proviruses are labeled with *, Syncytin-1 ORF is highlighted in bold. Underlined sequences retain an ORF without internal stop codons; italic sequences did not present frameshifts.

The obtained *Env* puteins were aligned and analyzed with respect to the Syncytin-1 amino acid sequence (NCBI reference NP_055405.3), showing a general accumulation of nucleotide substitutions, insertions and deletions. This led to the frequent occurrence of multiple premature internal stop codons and frameshifts, which prevents the effective production of a complete protein (Table 11 and Figure 15).

Noteworthy, seven *Env* puteins conserved a coding sequence without internal stop codons. Among them, three *env* genes (4q13.3, 5q11.2 and Xp22.31) are theoretically long enough to encode a complete protein. However, even if uninterrupted, those ORFs showed frameshifts with respect to the Syncytin-1 translation mode. 20q13.2 (483 aa) and 4q21.22 (320 aa) sequences are the most conserved with respect to Syncytin-1, presenting no stops and only one frameshift between positions 441-442 and 75-76, respectively. Xq22.3b (542 aa) and 9q22.31 (267 aa) present indeed no frameshifts but showed a single internal stop codon (position 39 and 149, respectively) that could potentially be reverted with a single point mutations, as already shown for Xq22.3b N-trenv [152].

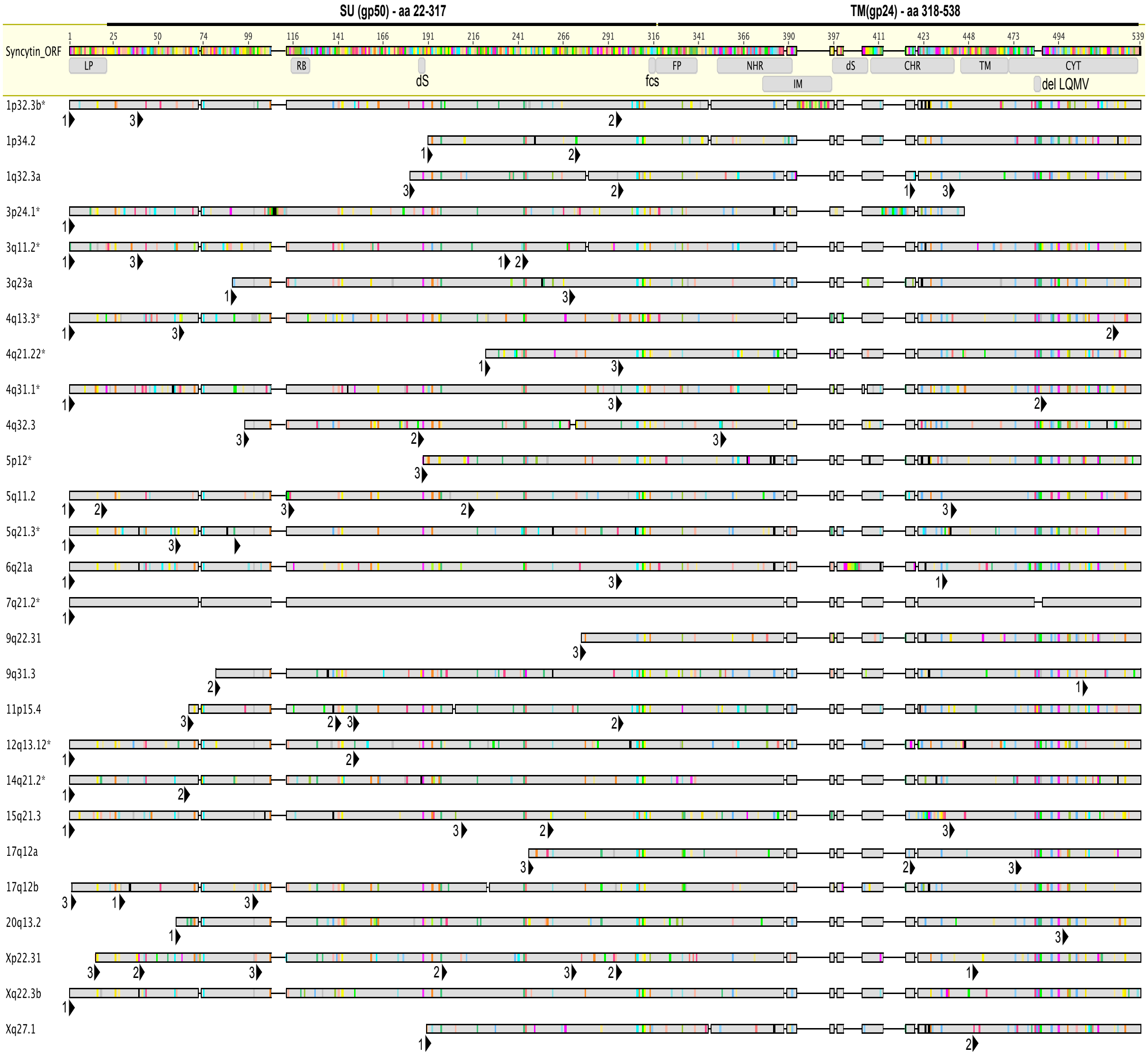


Figure 15. *Env puteins analysis with respect to Syncytin-1 Env protein (locus 7q21.2).*
In Syncytin-1 ORF, the domains involved in the protein structure and function are annotated: Leader Peptide (LP), Binding Domain motif (BD), SU and TM disulphide bounds motifs (dS), Furin cleavage site (FCS), fusion core N- and C- terminal Heptad Repeats (NHR and CHR), Immunosuppressive domain (IM), Transmembrane unit (TM), Intracytoplasmic Tail (CYT) and the relative Syncytin-1 specific deletion (LQMV del). In Env puteins amino acid substitutions and internal stop codons with respect to Syncytin-1 are labeled with colored and black lines, respectively. The reading frames are reported below each sequence with a number and an arrow.

Regarding the amino acid composition, as mentioned above all investigated *Env* puteins accumulated several substitutions, leading to a general average identity of about 85% with respect to Syncytin-1 sequence (Figure 15). To evaluate the puteins possible biological activity, we have characterized in detail the motifs known as involved in Syncytin-1 physiological function.

Primarily, the Env precursor must be processed into the mature SU and TM units, with a proteolytic cleavage that occurs at the Furin Cleavage Site conserved RKNR motif. The mutation of this domain has been reported to abrogate the proteolytic cleavage and the fusogenic activity of Env proteins, that exhibited also delayed kinetics of appearance on the membrane compared to the wild-type Env [30]. The RKNR motif of the HERV-W puteins was frequently mutated at the first position, mostly with the conversion of R residue to C or H (73% of analyzed ORFs), but was maintained in 7 sequences. After cleavage, SU and TM mature proteins are then linked through a covalent disulphide bond between the SU CWIC and the TM CX_6CC motifs. While the TM domain showed a high degree of amino acid homology with respect to Syncytin-1, in the SU motif we found an I>M substitution in 100% of sequences. Another fundamental step that drives the fusion activity is the interaction between the SU N-terminal 124 aa receptor binding domain and hASTC1 or hASTC2, which act as type D mammalian retrovirus receptors. In the binding domain, the $SDGGGX_2DX_2R$ motif, recognized as essential for the receptor contact, was retained in the 58% of sequences. The Syncytin-1 fusogenic activity is held by the TM portion, including a fusion peptide and a fusion core formed by the amino- and carboxy-terminal heptad repeats. In *Env* puteins the fusion peptide sequence was characterized by at least one substitution, with residue 332 (A) that was mutated in all sequences into an R or a G (and in one case into an E). Also the fusion core was affected by several mutations localized in both heptad regions, like the residue 433R>Q substitution that was present in 25 out of 26 carboxy-terminal repeats. Interestingly, the 75 amino acids long heptad repeat region showed a particularly high concentration of internal stop codons, harboring itself the 50% of the total stop codons found in the analyzed puteins. Moreover, in traditional Env proteins, the fusogenic activity is prevented by an inhibitory R peptide that is located in the TM intracytoplasmic tail and is normally removed by viral proteases. In Syncytin-1, a four amino acid deletion at the LQMV cleavage site made the protein constitutively competent for fusion [28], but this mutation was not present in any other analyzed HERV-W Env putein. Finally, the Syncytin-1 TM subunit also contains a conserved immunosuppressive domain that was thought to possibly contribute towards maternal immunotolerance [25, 49], even though other findings suggested the absence of this activity [48]. In any case, in the selected *Env* puteins, this domain presented several amino acid substitutions, showing also a premature termination at position 383 in 5 sequences. Hence, with respect to locus 7q21.2 Syncytin-1 protein, the other HERV-W loci *Env* puteins resulted highly defective, especially in sites involved in known physiological functions. However, despite these mutations, they may be still able to produce shorter proteins with biological significance, as observed for other HERV sequences [273].

Due to its maintenance despite the presence of huge recurrent flanking deletions affecting the 85% of HERV-W *env* genes, also the small *env* portion of about 30 nucleotides at position 8289-8318 was translated and compared with respect to Syncytin-1. All the 138 HERV-W elements that maintained this portion showed recurrent amino acid substitutions. In particular, the N in position 3 was changed in 136/138 sequences, substituted by H in 93% of the elements; while the V in position 8 was substituted in 135 sequences, showing a I in 90% of cases. This prevalence indicates that Syncytin-1 protein probably represents the exception, suggesting an unreported functional relevance of this short domain.

3.9 MSRV sequences homology with HERV-W elements

To complete the overview on the HERV-W group presence and impact on human genome, it was useful to consider also the proposed association with MS disease. In fact, the first HERV-W member was originally identified as cDNA sequences derived from particle-associated RNA in MS patients cultured cells [37, 38]. Those sequences were subsequently indicated as MSRV [22, 23, 111] and proposed to be an exogenous competent member of the HERV-W group, related to the MS development [113–115, 131]. Other reports, however, remarked the uncertain nature of MSRV [100, 116, 117] and proposed that some of these cDNA sequences could arise from the recombination of different HERV-W loci transcripts [151]. According to this hypothesis, such recombination could easily happen through RT switching templates, likely during *in vitro* PCR amplification, a common complication during the analyses of transcribed elements [13]. In particular, Laufer et al. proposed that 6 sequences previously published as MSRV elements could be traceable to a single HERV-W locus or to recombination events between two or more HERV-W loci transcripts [151]. Since other four sequences published as MSRV elements (accession numbers: AF009668, AF009666, AF009667 and AF123880) [22, 111] were not analyzed for possible HERV-W origin, having a more complete HERV-W database we analyzed them as described [151], including the 6 previously investigated MRSV sequences as internal control (Table 12).

Table 12. *HERV-W loci homology of previously described MSRV sequences and probes*

GenBank entry	HERV-W loci	Query cover	Discordant bases	Mapped portion in LTR17-HERV17-LTR17
AF127227 (544 bp)	3q23a* (99,5%)	1-544	3	env (8208-8752)
AF127228 (1932 bp)	Xq22.3b* (99,6%)	1-1932	9	pol-env (5444-5838 and 7682-9200)
AF127229 (2004 bp)	3p12.3* (99,9%) 18q21.32* (99,9%)	1-1084 1055-2004	2 2	pol-env-3′LTR (5452-6792 and 8290-8318 and 9115-9732)
AF123882 (2477 bp)	12q21.3* (99,8%)	1-2477	7	pol-env (5720-8199)
AF331500 (1629 bp)	Xq22.3b* (99,7%) 5p12* (99,4%)	1-1332 1308-1629	4 2	env (7720-9348)
AF123881 (1511 bp)	3q26.32* (99,9%)	1-1511	2	gag-pro (2765-4269)
AF009668 (2304 bp)	1p34.2 (99,1%) 2p12a (100%) 2p24.2 (100%) 6q27b (98,5%) 6q15 (97,2%) 3p12.3 (99,4%)	1-633 623-736 717-871 837-1424 1415-1763 1719-2304	6 0 0 11 10 4	pro-pol (4178-6480)
AF009666 (324 bp)	1p34.2 (99,5%)	1-324	3	pro-pol (4178-4521)
AF009667 (118 bp)	17q22 (98,2%)	1-118	2	pol (5031-5148)
AF123880 (1003 bp)	5p12 (99,6%) 3p24.1 (100%) 3q26.32 (98%)	1-203 198-593 592-1003	1 1 11	5′LTR (255-803)
AF072494 pol probe (678 bp)	6q21b (99,6%)	1-678	5	pol (4660-5338)
AF072496 gag probe (536 bp)	6q21b (99,6%)	1-536	2	pre gag-gag(2706-3199)
AF072497 pro probe (364 bp)	1p34.2 (99,2%)	1-364	4	pro-pol (4166-4522 and 5641-5549)
AF072498 env probe (591 bp)	Xq22.3b (99,5%)	1-591	3	env (8606-9196)

Previously published MSRV sequences and probes (column 1) were analyzed for their homology to one/more HERV-W locus/loci by BLAT search, considering the best match in human genome (reported in column 2 near to each HERV-W element). The MSRV elements portion similar to HERV-W locus/loci (column 3) as the number of discordant nucleotides with respect to the identified HERV-W locus/loci (column 4) and the correspondent positions in the LTR17-HERV17-LTR17 reference (column 5) were obtained through Mafft alignment and Geneious platform analysis. MSRV sequences were characterized through the analysis of each element with respect to the whole HERV-W dataset with Recco software.

* Already investigated by Laufer et al. 2009

° 95% similarity with AF135487, a retroviral-related sequence reported to be schizophrenia associated and mapped to multiple sites.

The analysis confirmed that AF127227, AF127228, AF123882, and AF123881 have high identity with a single HERV-W locus, while AF127229 and AF331500 could origin from recombination of two HERV-W loci transcripts. Similarly, AF009666 pro-pol and AF009667 pol mRNAs showed high identity with the HERV-W locus 1p34.2 (99,5% similarity) and 17q22 (98,2% similarity), respectively, while AF123880 5′LTR-pre gag mRNA showed sequence identity with three HERV-W loci (5p12, 3p24.1 and 3q26.32) with a similarity for each component ranging from 98% to 100%. Finally, AF009668 pro-pol mRNA showed a more complex identity pattern, with a high degree of mosaicism that seems to involve several HERV-W loci: 1p34.2, 2p12a, 2p24.2, 6q27b, 6q15 and 3p12.3. Interestingly, AF009668 shares a 95% similarity with AF135487, a retroviral-related sequence identified to be schizophrenia associated and also mapped to multiple sites [175].

Moreover, we performed the same analysis with the four MSRV DNA probes used to characterize the HERV-W placental expression. These probes were obtained through RT-PCR from RNA particles found in synoviocyte culture supernatants and pellets of a RA patient (AF072494 *pol*- and AF072498 *env*-probes respectively) and in B lymphocyte culture and plasma of a MS patient (AF072496 *gag*- and AF072497 *pro*-probes respectively) [23]. Both AF072494 *pol*- and AF072496 *gag*-probes showed high identity with HERV-W 6q21b locus (99,6% similarity), while AF072497 *pro*- and AF072498 *env*-probes were highly identical to locus 1p34.2 (99,2%) and Xq22.3b (99,5%), respectively.

In the light of a suggested immunopathogenic role of MSRV Env proteins, the MSRV sequences containing an *env* gene (or a portion of it) and showing highest identity with one of the HERV-W loci analyzed for *Env* puteins were manually translated and aligned with the correspondent HERV-W *Env* putein and the Syncytin-1 protein for further comparison (Figure 16). Interestingly, with respect to the Syncytin-1 sequence, the HERV-W and the correspondent MSRV puteins shared the great majority of amino acid substitutions, and often the same amino acid change was common to all sequences analyzed. In particular, AF127227 and 3q23a share the same frameshift at position 270 of Syncytin-1 sequence, and AF127227 and AF127228 showed an internal stop codon at the same position observed in 3q23a and Xq22.3b, respectively (position 39, W in Syncytin-1). Differently, AF331500 lacks this internal stop codon, presenting, like Syncytin-1, a W in the correspondent position. As already observed for HERV-W, also MSRV *Env* puteins showed at least one amino acid change in all domains relevant to Syncytin-1 biological activity. Given the proposed MSRV Env proteins role in pathogenesis, the presence of shared recurrent substitutions, possibly preventing the MSRV *Env* puteins functionality as compared to Syncytin-1, opens further questions that will have to be addressed. Overall, while more MSRV RNA expression studies are needed, the here reported HERV-W genomic map and characterization is a further step to properly assess the MSRV/HERV-W role in the context of MS.

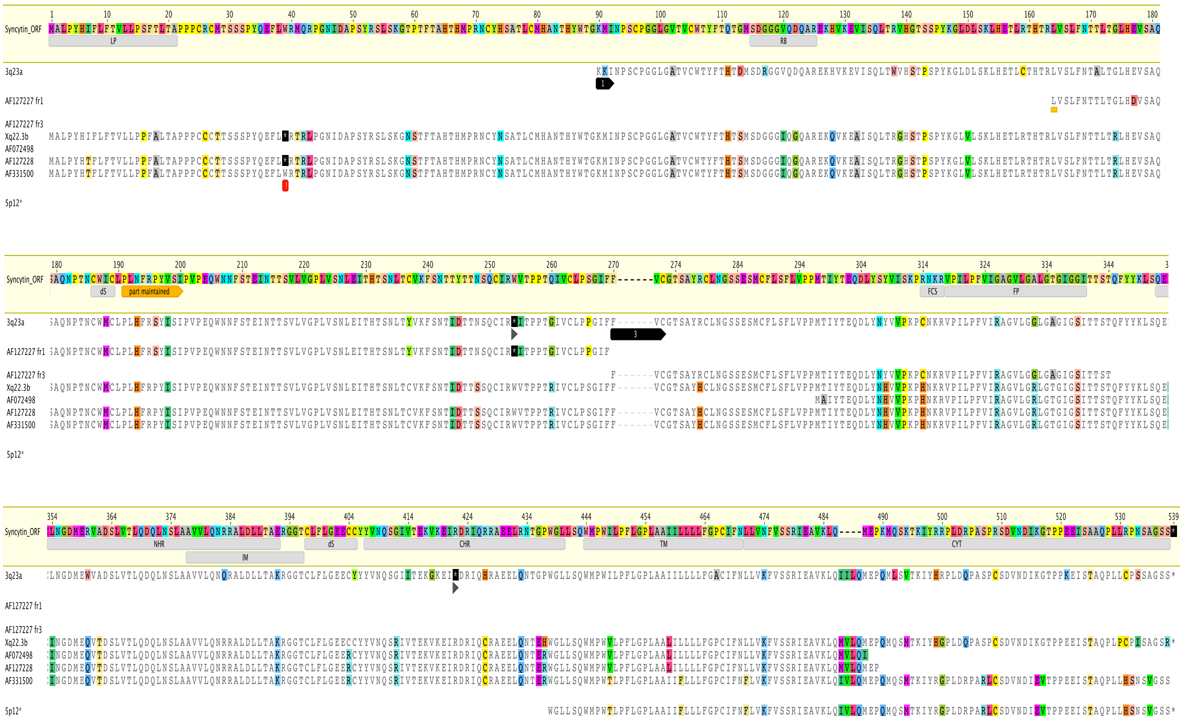


Figure 16. *MSRV Env puteins analysis with respect to the most similar HERV-W loci Env puteins and Syncytin-1 (locus 7q21.2).*
In Syncytin-1 ORF, the domains mostly involved in the protein structure and function are annotated: Leader Peptide (LP), Binding Domain motif (BD), SU and TM disulphide bounds motifs (dS), Furin cleavage site (FCS), fusion core N- and C-terminal Heptad Repeats (NHR and CHR), Immunosuppressive domain (IM), Transmembrane unit (TM), Intracytoplasmic Tail (CYT) and the relative Syncytin-1 specific deletion (LQMV del). In MSRV and HERV-W Env puteins, amino acid substitutions and internal stop codons with respect to Syncytin-1 are labeled with colored and black squares, respectively. The reading frames are reported below each sequence with a number and an arrow. The orange arrow indicates the Env portion frequently maintained in presence of flanking huge recurrent deletions.

3.11 Discussion

Since the discovery of Syncytin-1 role in placentation [23–25, 274], a great attention has been dedicated to the expression potential of the HERV-W group, trying to further understand its impact on the host. Many studies were focused on HERV-W correlations with several human diseases, primarily represented by MS [16–19, 28, 75, and reviewed in 76] and other major neurological pathologies such as schizophrenia and bipolar disorder [181, 185, 275]. Despite this broad investigation, no certain correlations between HERV-W group expression and any human disease has been confirmed, and even in the major field of MS the findings are still highly discordant [100]. One of the problems faced in this scenario is the unfortunate lack of a complete and updated description of the HERV-W sequences in the human genome, including a detailed characterization of HERV-W single members and their genomic background. Such information could help in better interpreting the wide range of HERV-W expression data in both physiological and pathological contexts.

Therefore, using more updated genome data and a double bioinformatics identification approach, we analyzed GRCh37/hg19 assembly identifying a total of 213 HERV-W unambiguously classified members. Each HERV-W sequence has been precisely localized and characterized in term of structure, phylogeny and evolution, allowing to specifically identify the uniqueness of each HERW-W single member, and highlighting a number of non-previously reported characteristics of the group.

Firstly, we observed several nucleotide differences in HERV-W members with respect to the assembled LTR17-HERV17-LTR17 reference, which was built on a small number of sequences and therefore does not properly represent the entire group. Secondly, we classified the HERV-W sequences into two subgroups through a LTRs phylogenetic analysis strongly supported by the identification of key mutated positions in both LTRs, shared by the majority (from 95% up to 100%) of sequences within the same subgroup. Beside LTRs mutations relevant for classification purposes, the subgroups comparisons showed single nucleotides differences along the whole retroviral sequence. For this reason we propose three new consensuses, one general HERV-W proviral consensus (HERV-W_PV-consensus) and one consensus for each subgroup (HERV-W_SG1 and HERV-W_SG2 consensus) (Additional file 8 in [21]), that, in our opinion, better represent the overall HERV-W group composition.

In the present study, for the first time, the period of insertion has been estimated for each HERV-W locus through at least two different methods of age calculation. This provided a precise and exhaustive picture of the group diffusion among primates, and brought important improvements in the method reliability and applicability. Moreover, the analysis showed significantly different dynamics in the two subgroups diffusion, pointed out also by the analysis of the PBS type variability.

The analysis of structural features previously described for Gammaretroviruses [257] in the HERV-W single members allowed to characterize them for the first time in term of prevalence and sequence conservation among the group. Noteworthy, in addition to the traditional Gag NC Zinc finger motif [276], we found a previously unreported second Zinc finger with an unusual structure, lacking one variable residue. Another interesting feature reported here for the first time is the presence of a weak bias in the HERV-W elements purine amount, with enrichment in A and a consequent underrepresentation of G.

With regards to the group genomic context of integration, we provide an updated overview of the 80 HERV-W elements inserted into human genes, and of the HERV-W sequences predicted capacity to bind cellular TFs. In particular, 55 HERV-Ws were found into coding genes, 8 more than what previously observed [119, 268], while 25 elements were inserted in human non-coding genes, of which the great majority (22) are reported here for the first time. *Env* putein analysis led us to identify and functionally characterize 16 full-length or near full-length *env* genes, 3 more than previously reported [272], and 10 conserved but shorter *env* genes. Although the relative puteins resulted highly defective and mutated in comparison to Syncytin-1, these genes may still be able to produce shorter proteins with a biological significance, as observed for other HERV sequences [273].

In the light of the debated connection between HERV-W loci expression and MS disease, we investigated the elements known as MSRV in order to evaluate their identity with respect to one or more HERV-W loci, in agreement to what has been previously reported [151]. Our results confirmed that the majority of MSRV related sequences has 97% to 100% identity with one single HERV-W locus, but more complex pattern of identity, apparently involving 3 or even 6 loci, were also observed. Furthermore, the comparison between MSRV *Env* puteins and the most identical HERV-W puteins showed common amino acid substitutions with respect to Syncytin-1, affecting all domains relevant for its activity.

In conclusion, this report provides, to our knowledge, the most exhaustive and updated HERV-W group description up to date in terms of structure, evolution and context of integration into the human genome, revealing that this polymorphic multicopy family is not only represented by the single HERV-W member Syncytin-1. We showed that HERV-W elements were acquired by primates during a rather long time period, and evolved within and with their genome that exerted a selective pressure leading to the modification of HERV-W structures, including the previously shown co-option of one member for an important physiological function [24, 25]. Overall, the here presented characterization of the HERV-W composition and genomic context of insertion will be essential to investigate the effects that, beside protein expression, HERV-W sequences can exert in the different human tissues, both under physiological conditions and regarding the putative involvement in human disease etiology, to finally define their real impact and contribution to our genome.

Chapter 4. HERV-W group evolutionary history in non-human primates: characterization of ERV-W orthologs in Catarrhini and related ERV groups in Platyrrhini

4.1 Introduction

The genome of all vertebrates includes a portion of sequences of viral origin, named Endogenous Retroviruses (ERVs), which are remnants of infections occurred mostly along the last 100 million years [277]. While in some vertebrates ERVs and their exogenous counterparts are currently coexisting [278–280], the exogenous retroviruses that originated human ERV (HERV) insertions have gone extinct, mostly million years ago (MYa), and cannot thus provide information on their origin and evolution. In light of this, the nearest information about HERVs original characteristics could be found through the comparison of the same (orthologous) elements within the evolutionarily related species that share them.

The human genome is composed of 8% of HERV sequences [1], recently classified into 39 "canonical" and 31 additional "non canonical" groups [8]. A number of these HERV groups have been acquired by primates before the separation of *Catarrhini* (which includes the family *Cercopithecidae*, also known as Old World Monkeys, OWM, and *Hominoidea*) and *Platyrrhini* (also known as New World Monkeys, NWM) parvorders, that occurred ~ 40 million years ago (MYa) [221, 222]. These ancient groups are thus shared between *Catarrhini* OWM and *Platyrrhini*, such as in the case of HERV-L and HERV-H sequences [281]. Many other HERV groups, such as HERV-E and HERV-K(HML2), are indeed younger, having been acquired after *Catarrhini* and *Platyrrhini* evolutionarily separation. This means that these groups entered the primates lineage later than 40 MYa, being thus shared by OWM and hominids, but resulting absent in NWM.

During the last decades, HERVs acquired a growing importance due to the discovery of relevant physiological functions, and in general for their potential effects on the host [238, 282, 283]. HERV sequences can, in fact, affect host genome and genes in various ways. For instance, HERVs can provide gene-regulatory functions [55, 235, 238, 239, 247, 284–287] and cause genetic alterations, such as insertional mutagenesis and gene disruption [2, 234–236, 238, 277]. Pathological roles have been suggested for some HERVs [238, 282, 283] and HERV expression has been tentatively linked to a number of human diseases [20, 74, 90, 91, 288], yet no unequivocal cause-effect relationships have been established so far [20, 117, 238]. In particular, the HERV-W group has drawn considerable interest because of important physiological functions and proposed involvement in human diseases.

In the previous chapter, we identified and described in detail the distribution and genetic composition of 213 HERV-W loci present in the human genome assembly GRCh37/hg19, providing an updated and complete overview of the group [21]. Just to summarize, the HERV-W group comprises 65 proviruses (acquired with traditional retroviral infection and showing complete LTRs), 135 processed pseudogenes generated by the LINE retrotransposition machinery [33, 289], and showing typically truncated LTRs [33] and 13 undefined elements, which lack both LTRs and could not be subsequently assigned to the two classes. Moreover, the phylogenetic and structural analysis led to the classification of HERV-W members into two subgroups, named 1 and 2, that, based on time of integration estimation, were acquired in the primates lineage approximately between 40 and 20 MYa, after the divergence between *Platyrrhini* and *Catarrhini* [21].

In order to further characterize the HERV-W group origin and evolution, we investigated its presence among the primates lineage leading to humans, with a deeper focus on primate species with publicly available genome sequence assemblies (Figure 17).

In particular, we i) collected and analyzed the ERV-W loci orthologous to the previously characterized HERV-W sequences in the genome sequences of 5 primate species belonging to *Catarrhini* parvorder, specifically Rhesus Macaque OWM and *Hominoidea* great apes (Gibbon, Orangutan, Gorilla and Chimpanzee), that evolutionarily diverged ~ 30 MYa; ii) identified hitherto uncharacterized ERV elements closely related to ERV-W in *Platyrrhini* species Marmoset and Squirrel Monkey (*Cebidae* family), possibly suggesting the presence of a common ancestor of those sequences, previously named ERV1-1 in RepBase, and the HERV-W group; iii) analyzed NCBI Trace Archive unassembled genome sequences data, providing support for the presence of such ERV-W closely related elements also in two species belonging to the other *Platyrrhini* families, namely *Atelidae* and *Pitheciidae*; and iv) corroborated the lack of (H)ERV-W closely related elements in the most primitive *Tarsiiformes* and in *Prosimians* (including *Lemuriformes* and *Lorisiformes*).

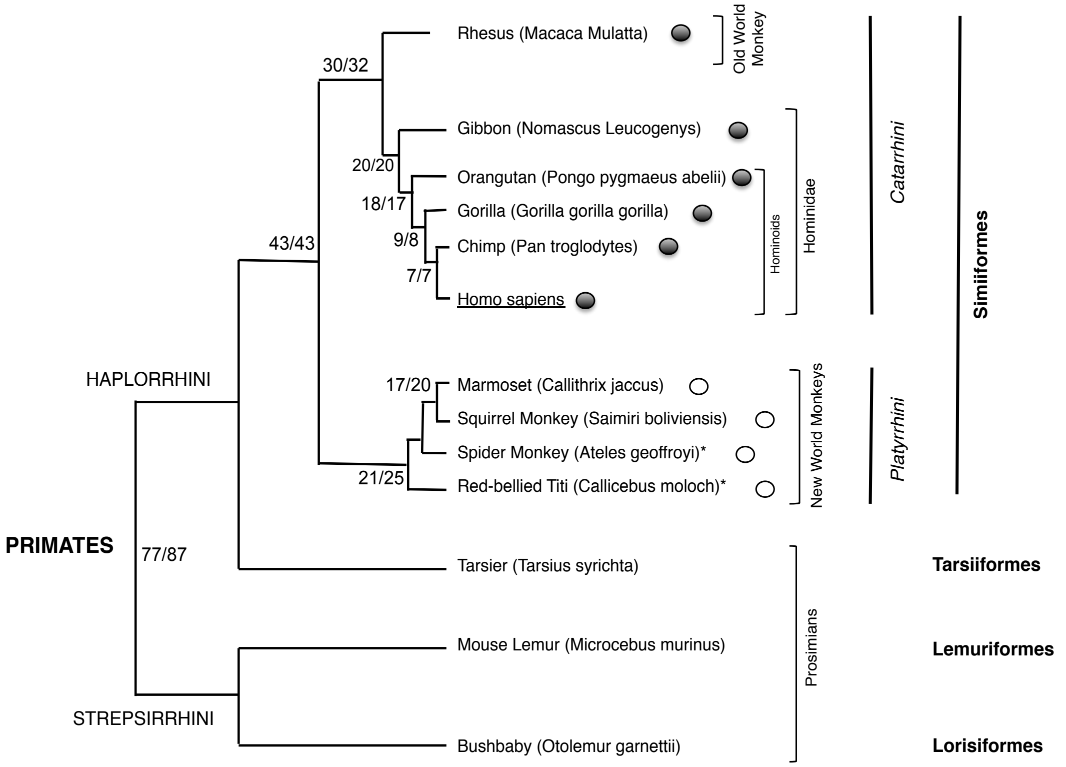

Rhesus (Macaca Mulatta)
Old World Monkey
30/32
Gibbon (Nomascus Leucogenys)
20/20
Orangutan (Pongo pygmaeus abelii)
18/17
Gorilla (Gorilla gorilla gorilla)
9/8
Chimp (Pan troglodytes)
7/7
Homo sapiens
43/43
Hominoids
Hominidae
Catarrhini
Simiiformes
HAPLORRHINI
17/20
Marmoset (Callithrix jaccus)
Squirrel Monkey (Saimiri boliviensis)
Spider Monkey (Ateles geoffroyi)*
21/25
Red-bellied Titi (Callicebus moloch)*
New World Monkeys
Platyrrhini
PRIMATES
77/87
Tarsier (Tarsius syrichta)
Tarsiiformes
Mouse Lemur (Microcebus murinus)
Lemuriformes
Prosimians
STREPSIRRHINI
Bushbaby (Otolemur garnettii)
Lorisiformes

***Figure 17**. Schematic representation of the phylogeny of the primate species analyzed in this study.*
Presence of (H)ERV-W sequences in respective species is indicated with a filled circle. Presence of (H)ERV-W closely related elements with an empty circle. Parvorders and infraorders of primates are indicated in italics and bold, respectively, letters. Estimated ages of divergences of evolutionary lineages are given, in millions of years ago, near tree nodes. Ages were taken from Steiper and Young 2006 (first numbers) and Perelman et al 2011 (second numbers). Primate species marked with an asterisk lack assembled reference genome sequences.

Taken together, the data obtained provide a more clear and detailed analysis of the HERV-W group presence and distribution within primates genomes, and further depict the evolutionary history of ERV-W in various primate lineages. Importantly, the comparative analysis allowed to characterize ERV-W species-specific insertion among *Catarrhini* primates, further detailing the group period and dynamics of colonization. Finally, the individuation of hitherto unreported ERV-W related elements in *Platyrrhini* species provided important insights into putative ancestral sequence contributions to ERV-W.

4.2 Comparative analysis of the HERV-W orthologous sequences in Catarrhini primates

Following the characterization of a total of 213 HERV-W sequences inserted into the human genome assembly hg19 (see ref. [21] and Chapter 3), we analyzed in more detail the presence or absence of orthologous loci of human HERV-W loci in the genome sequences of non-human primate species. For the sake of simplicity, we will refer to the respective non-human primate sequences as ERV-W, in order to distinguish them from the human (HERV-W) sequences. Making use of homologous genome regions and annotations provided by the UCSC Genome Browser [214, 290, 291], the presence of HERV-W orthologous loci was examined for the genome sequences of OWM species Rhesus Macaque and *Hominoidea* species Gibbon, Orangutan, Gorilla and Chimpanzee by comparisons of respective genomic loci. To properly verify the presence of each ERV-W locus, due to the high nucleotide identity shared by the various ERV-W copies, we also put special emphasis on sequence similarities in genomic regions immediately flanking, both up- and down-stream, a regarded ERV-W insertion site. Of note, since no comparable sequence information was available for the 2 HERV-W loci on chromosome Y, except for Chimpanzee, we included in the analysis the remaining 211 HERV-W loci. Our investigation thus generated a relatively exhaustive comparative map of orthologous ERV-W locus insertions in primate genomes (Table 13). *Hominoidea* species Chimpanzee, Gorilla and Orangutan genome sequences revealed a number of orthologous ERV-W loci overall comparable to the number found in human genome assembly GRCh37/hg19 [21], with a total of 205, 207 and 205 loci, respectively, representing orthologs to human loci (Table 14). Gibbon and Rhesus genome sequences harbored 190 and 131 human-orthologous ERV-W loci, respectively. In some instances, the absence of an entire ERV-W insertion in some primates could be due to, on the one side, its integration into the primate genome after the separation of the derived evolutionary lineages, and thus providing direct information about the time period of germ line colonization of the group. On the other side, the lack of ERV-W loci could be due to subsequent deletions or other rearrangements of genome regions, errors in genome sequence assemblies, or errors in comparative analysis of genome sequences, particularly for primate species with less complete genome sequences.

Based on our analysis, 123 out of 211 HERV-W loci as orthologous loci being shared by all *Catarrhini* primates analyzed. When considering in this analysis only the (H)ERV-W loci present in Rhesus and humans, excluding hence those ERV-W loci apparently absent in some intermediate primates because of potential genome sequence issues (see above), the number of ERV-W loci shared from Rhesus to humans increases to 131/211 (Figure 18). Those findings further corroborate that the first and major wave of ERV-W locus formations occurred in the primate lineage between 43 to 30 MYa, after separation of *Catarrhini* and *Platyrrhini,* but before divergence of Rhesus OWM from *Hominoidea*, in line with previously reported time periods of integration [21, 34, 256].

Table 13. *Comparative analysis of HERV-W loci in the human reference genome sequence and ERV-W loci in reference genome sequences of non-human Catarrhini primates*

Human Locus	Chimpanzee	Gorilla	Orangutan	Gibbon	Rhesus
1p34.2	1:42253290-42258830	1:43411475-43416951	1:188148897-188152813	12:1199243-11200624	1:45086710-45090786
1p33a	1:46756770-46762318	1: 48028689-48035081	1:183499440-183502466	12:15590272-15596434	x
1p33b	1:47324219-47330023	1: 48581048-48586853	1:182873207-182877475	12:15989128-15990596	1: 50318172-50331476
1p32.3a	1:51601144-51605252	1:52885357-52887891	1:178547060-178549590	12:20054000-20057336	1: 54515446-54519458
1p32.3b	1:55408053-55411656	1:56642603-56648849	1:174801151-174802301	5:35333628-35335397	1:58267573-58268352
1p32.2	1:56284257-56289559	1:57530345-57536033	1:173918404-173920056	5:36161801-36162549	1:59137876-59143907
1p22.2a	1:89720008-89727789	1:91355585-91361401	1:139918001-139925607	12:87637049-87637859	1:92353028-92361547
1p22.2b	1:91947016-91949019	1:93563999-93565452	1:137658234-137660239	12:85633074-85635095	1:94612649-94615849
1p13.3	1:126991829-126997725	x	1:118421928-118423397	12:67891509-67898817	x
1p12	1:117540334-117541092	1:122336233-122339314	1:108900732-108903790	12:58509513-58512578	1: 122805532-122811574
1q22	1:133880582-133884859	1:134614317-134618627	1:95886163-95889889	x	x
1q25.2	1:156663510-156665355	1:157566335-157568179	1:72735143-72736984	12:24182099-24183924	1:208448557-208450365
1q32.1	1:184733125-184738291	x	1: 44308859-44314242	1:55847121-55852084	1:165106368-165111778
1q32.3a	1:190976937-190979014	1:191923501-191925579	1:38129378-38131503	5:50109609-50110550	1:159094276-159096392
1q32.3b	1:190979015-190980062	1:191925580-191926627	1:38128333-38129377	x	1:159093063-159093648
1q42.13	x	x	1:21962650-21964249	x	1:143310829-143314813
2p25.3	2A:144467-152114	2A:152261-164890	2a:112878839-112887231	19:138696-138806	13:129815-130665
2p24.2	2A:17538180-17545961	2A:17658411-17663167	2a:95032893-95042955	19:17047420-17055204	x
2p23.1a	2A:30942691-30945867	2A:31075188-31078395	2a:81542854-81548617	19:59744922-59751046	13:28737325-28743099
2p23.1b	2A:-32086090-32091511	2A:32214122-32219110	2a:80402060-80419150	x	x
2p22.3	2A:34160122-34164672	2A:34291350-34298765	2a:78320607-78323456	19:56637441-56643813	13:31843251-31850558
2p16.2	2A:54602365-54607165	2A:54725292-54731778	2a:57295362-57300173	14:21746828-21751867	13:53860183-53864925
2p12a	2A:76874262-76881802	2A:77236237-77242452	2a:34727369-34731021	14:61264724-61272263	x
2p12b	2A:80223274-80230377	2A:80587451-80594625	2a:31313548-31314408	14:64256008-64263165	13: 79594807-79601741
2q11.2	2A:96457998-96464207	2A:93693340-93699553	2a:4351994-4358635	14:47551778-47552561	x
2q12.2	2A:106558520-106564927	2A:103822886-103829321	2a:14106781-14114101	14:79309923-79312090	13:105917912-105922876
2q13	2A:112164165-112169801	2A:109841633-109847228	2a:20003401-20008321	14:75631185-75634414	13:111233128-111238804
2q22.1	2B:140167036-140168508	2B:23312386-23313859	2b:25164017-25165488	20:15952806-15954277	x
2q22.2	2B:146950391-146958340	2B:30135636-30144016	2b:32152722-32160817	20:9208892-9209164	12:5812408-5824356
2q22.3	2B:151071967-151074566	2B:34323965-34326540	2b:36426850-36429410	20:5095433-5098009	12:9947298-9950263
2q24.3	2B:168960094-168961800	2B:52345755-52347454	2b:54547038-54548733	17:13461594-13463294	x
2q31.1a	2B:175999149-176001226	2B:59399792-59402420	2b:61653777-61655846	22a:61412377-61414448	x
2q31.1b	2B:179783615-179784879	2B:63242466-63243732	2b:65455060-65456328	22a:65278805-65280074	x
2q31.2a	2B:181901536-181905470	2B:65377456-65381387	2b:67606265-67610287	22a:67506007-67509944	12:40727453-40731380
2q31.2b	2B:182953883-182955007	2B:66488533-66489670	2b:68700743-68701881	22a:68568092-68569234	x
2q31.3	2B:184742283-184749930	2B:68304396-68312032	2b:70583897-70591501	22a:70391297-70392082	12:43586070-43586837
2q32.3	2B:199988015-199993634	2B:83666472-83671864	2b:86236960-86242377	22a:85408131-85413538	12:58745222- 58750776
2q35	2B:222640652-222642576	2B:106505250-106507174	2b:109413971-109415910	22a:108478783-108480701	12:81269029-81271117
2q37.3	2B:244279468-244281196	2B:128414265-128416317	2b:131577186-131578236	22a:130089703-130091447	12: 102807079-102809155
3p24.3	3:20173751-20182863	3:20467991-20470036	3:127597549-127606096	4:133054750-133061805	x
3p24.1	3:27424905-27430298	3:27818251-27822673	3:120175044-120180084	4:125712312-125717730	x
3p22.2	3:38875248-38883128	3:39334776-39342677	3:108404235-108412032	4:114158715-114166510	2:99671284-99680777

3p22.1	3:40286651-40288901	3:40732623-40734881	3:107079765-107082032	4:112819282-112821521	2:98230660-98231219
3p21.31	3:49075534-49081244	3:49571043-49576764	3:98224922-98230630	x	2:89468308- 89473630
3p12.3	3:76232812-76238047	3:76689796-76697458	3:70669488-70674769	21:68527020-68531348	2:62781497-62786415
3p11.1	3:89553984-89559542	3:90119008-90124604	3:55076036-55090089	21:31520654-31521475	x
3q11.2	3:99721235-99729919	3:95667711-95676260	3:36624812-36633451	21:24951253-24952047	x
3q13.31	3:119100608-119106454	3:115196654-115204808	3:17000577-17002788	21:5864030-5866502	2:3583024-35832590
3q13.32	3:122220200-122228424	3:118345276-118353836	3:13831917-13839900	21:2640450-2647105	2:38943952-38945205
3q22.1	3:134666547-134672009	3:131258946-131261155	3:133508307-133510533	x	x
3q22.2	3:139461628-139465114	3:136074458-136077379	3:138385499-138389037	8:15133779-15137331	x
3q23a	3:145371784-145373539	3:142084973-142086733	3:144355084-144356823	x	x
3q23b	3:146014559-146019921	3:142726670-142734856	3:144983797-144989126	8:8536966-8542355	x
3q25.1a	3:153412804-153415159	3:150221123-150223482	3:152575847-152578207	11:74063566-74065922	2:140536212-140538676
3q25.1b	3:154541067-154547814	3:151330936-151338003	3:153751053-153758375	11:75242510-75248141	2:139390931-139396793
3q25.2	3:158711238-158712463	3:155648022-155649256	3:158106513-158107738	11:79487320-79488523	x
3q26.1a	3:166111212-166116905	3:163162512-163167502	x	x	x
3q26.1b	3:167473934-167478504	3:164516981-164521565	3:167191696-167196289	11:88175752-88180619	2:126200275-126201326
3q26.31	3:176563297-176564859	3:173763132-173764516	3:176304955-176306426	11:97192481-97194326	2:116791875-116792862
3q26.32	3:183023551-183028246	3:180256330 180261570	3:182666427-182671678	11:103740269-103745536	2:109856451-109863085
3q28	3:195751327-195757290	3:193307768-193316413	3:19621452-19623134	11:116572370-116579114	x
4p16.3	4:205053-210209	4:178027-183271	4:179190-184674	x	1:3830508-3835584
4p16.1	4_GL389982_random:106478-114099	4:8873503-8880956	4:4667520-4677688	20:81106451-81111153	5:541690-544951
4p15.1	4:33683797-33685959	4:33947794-33949620	4:34513514-34515578	20:53240062-53241905	x
4p14	4:36412069-36415535	4:36663214-36666669	4:37385204-37388628	20:51006391-51009806	5:31771891-31775321
4p13	4:42359552-42366991	4:42699519-42707323	4:43383829-43393863	20:45326767-45327474	5:37698404-37700405
4q13.1	4:66923998-66929853	4:71238204-71243766	4:56865343-56868963	9_JH996053_1_random: 26962-27735	5:67411853-67414680
4q13.3	4:57165696-57173240	4:81447093-81459532	4:76151524-76159062	9:57497999-57508351	x
4q21.22	4:47539907-47546317	4:91281843-91282693	4:85981186-85990533	9:67482787-67487623	5:47444991-47451761
4q21.23	4:87606550-87609269	4:94263581-94265177	4:89135103-89137819	9:70354069-70356817	5:79310265-79312983
4q23	4:101549701-101554898	4:108280955-108286174	4:103377721-103383317	9:83889975-83895189	x
4q24	4:108063483-108064522	4:114821658-114822698	4:110162500-110163552	9:90312786-90313839	5:99786345-99788006
4q25	4:112903618-112909625	4:119677485-119684465	4:115006746-115012353	18:60416563-60422134	5:104468970-104475494
4q26	4:116737929-116745413	4:123552155-123559161	4:118783654-118791052	18:64303064-64310010	x
4q28.3	4:135488458-135495431	4:142804521-142811390	4:138034034-138040898	7b:23751814-23758537	5:126636585-126643134
4q31.1	4:141331404-141336634	4:141331404-141336634	4:144075670-144081131	7b:18158248-18168408	5:132509747-132519454
4q31.3	4:155881471-155881974	4:163112758-163113261	4:158588117-158588619	7b:76717873-76718371	5:146548931-146549414
4q32.3	4:167758388-167760911	4:175172704-175175230	4:171074691-171077217	7b:88505785-88508323	5:158381124-158383467
4q33	4:173389541-173397441	4:180773396-180780266	4:176886113-176892928	7b:94029987-94037088	5:163791193-163798967
4q35.1	4:186323457-186328979	4:193792460-193797151	4:191171387-191176663	7b:107113762-107118593	x
5p13.3	5:83846964-83849602	x	5:32505792-32507178	6:30447198-30448571	x
5p13.2	5:78819749-78824172	17:57384319-57388734	5:37264492-37268896	6:34889121-34894348	6:36579091-36583756
5p12	5:70913580-70917565	17:49504788-49508734	x	7b:40701857-40703881	x
5q11.2	5:57867187-57870060	17:39633141-39636012	5:58842785-58845644	18:69006413-69008913	6:55582968-55585837
5q12.1	5:54661119-54669394	17:36369435-36373496	5:90947585-90954053	18: 72207728-72215689	6:58785178-58785613
5q14.3a	5:27266556-27267659	5:70983137-70984249	5:88711121-88712233	2:102323720-102324844	x
5q14.3b	5:25629324-25630237	5:72610586-72611738	5:90380302-90381452	2:103869732-103870885	x
5q21.3	5:108984997-108985749	5:91520180-91520605	5_random:12862196-12862975	2:122626095-122626789	x
5q22.2	5:112909771-112917701	5:95450247-95458049	5:113262177-113262992	x	x

6p25.3	6:1292049-1293617	6:1261365-1262966	6:1284503-1286313	x	x
6p23	6:13981483-13986911	6:14489219-14494833	6:14389231-14394316	8:64669373-64673653	4:13918950-13923675
6p22.3	6:24931739-24938036	6:25545916-25552550	6:25751300-25757034	8:75670082-75674265	x
6p12.2	6:53641965-53642979	6:54374437-54376535	6:53289866-53291870	22A:51653820-51659432	4:53171292-53172040
6q12	6:65069311-65076302	6:64581749-64587223	6:65268103-65275163	3:66535862-66543234	x
6q14.1	6:81863564-81867585	6:81290133-81294639	6:82110915-82112102	18:4086703-4087476	x
6q14.2	6:84005313-84010667	6:83426045-83432065	6:84307325-84313345	18:1909813-1915353	4:81425739-81431773
6q14.3a	6:85288618-85298772	6:84715543-84725125	6_random:5962213-5967827	18:612954-622397	4:82672432-82681262
6q14.3b	6:85521241-85522334	6:84955542-84956633	6:85854260-85855109	x	x
6q15	6:89039681-89047267	6:88486061-88491508	6:89336036-89343758	3:75135111-75142228	x
6q21a	6:107251202-107258843	6:106280485-106289388	6:108164419-108169412	3:93527711-93533503	4:104450896-104458619
6q21b	6:108360769-108362102	6:107266958-107268293	6:109303096-109304431	3:94489932-94491265	4:105499863-105501075
6q21c	6:112212558-112221064	6:111267599-111277982	6:113312820-113319850	3:98341512-98350979	4:155062305-155068949
6q23.3	6:139077927-139081234	6:138471628-138475067	6: 140567533-140570958	3:125091871-125095307	4:128358807-128359217
6q24.2a	6:144537244-144545134	6:143915791-143923659	6:146011206-146019291	3:130476441-130484982	4:122831505-122836573
6q24.2b	6:145594905-145600164	6:145011688-145016929	6:147063019-147068230	3:131522318-131527635	4:121778282-121782924
6q27a	6:168196050-168198415	6:167660421-167662777	6:169977769-169980110	3:153932588-153934961	4:165485378-165485777
6q27b	6:168779571-168786554	6:168245075-168252238	6:170602873-170612742	3:154543438-154550502	4:166076882-166083904
7p21.3	7:11454683-11455415	7:12946790-12947523	7:71926836-71927568	11:35202817-35203548	3:113921735-113922135
7p21.1	7:15260008-15260819	7:16791296-16792122	7:68014501-68015326	11:39154170-39154995	3:110036511-110037343
7p14.2	7:34258374-34259253	7:36333318-36334197	7:48352638-48353514	17:72913054-72913918	3:91006143-91007026
7p14.1	7:42012834-42013613	7:40970422-40973177	7:43975324-43981639	17:68443538-68446215	3:86780917-86787858
7q21.2	7:92994461-93004661	7:89728903-89739105	7:83671198-83681292	11:23622600-23632779	3:125395609-125401290
7q31.1a	7:109830091-109837285	7:106027213-106034405	7:104537519-104544698	13:59332727-59339920	3:147229086-147235613
7q31.1b	7:115853464-115860512	7:112259759-112265586	7:110874280-110881545	13:65405359-65406334	x
7q31.31	7:-121070951-121075564	7:117512020-117516592	7:116330386-116335005	13:70797033-70801646	x
7q31.32	7:123707562-123708067	7:120183202-120189233	7:119043331-119051911	13:73467037-73467434	3:161207354-161207846
7q32.3	7:133314031-133318098	7:129962376-129963726	7:128806239-128807453	13:82934202-82938260	3:170800071-170805127
7q33	7:136067062-136074671	7:132789695-132797295	7:131682479-131689985	13:85802943-85810470	3:173518492-173526879
7q36.1	7:150944189-150949844	7:148113756-148129415	7:147466452-147472217	13:100786358-100792072	3:188993128-189004050
8p21.3	8:-16281266-16286506	8:19938116-19943036	8:19676983-19681997	8:25630481-25635373	8:20087689-20091694
8q11.21	8:-45885296-45888733	8:46024639-46028122	8:49731551-49733542	x	x
8q12.1	8:58148573-58155404	8:58431245-58438387	8:62439349-62445448	16:51136942-51143099	8:64784201-64790544
8q12.3a	8:60410570-60412593	8:60736145-60737231	8:64681558-64683562	16:9521148-9521468	x
8q12.3b	8:62578005-62583417	8:62927255-62932694	8:66986512-66991875	16:7378490-7383896	8:69189857-69195271
8q13.2	8:65783643-65785306	8:66223413-66225075	8:70146769-70151335	16:4221067-4222732	8:72353987-72355647
8q21.11	8:71793253-71793992	8:72222477-72223230	8:76299655-76300393	16:34873117-34873854	x
8q21.13	8:78804030-78807285	8:79281016-79284491	8:83624756-83628418	16:27878400-27881858	8:85499224-85502665
8q24.13	8:123575956-123584217	8:124658894-124664757	8:132594312-132604943	16:79856860-79857629	8:129858564-129864063
9p24.1	9:8839731-8841033	9:8927315-8928614	9:54013050-54014349	1a:9724599-9725900	x
9p21.3	9:23148987-23149797	9:23263632-23264473	9:39126285-39129266	8:90030005-90030819	15:55124164-55124977
9p21.1	9:30014930-30017487	9:30195780-30198317	9:32010647-32010847	1a:71021669-71024163	15:48317658-48320188
9p13.3	9:36074441-36076965	9:36374651-36377174	9:25917763-25920272	8:82230477-82233512	x
9q22.1	9:87708804-87713235	9:70849219-70853676	9:84393498-84398032	1a:46130234-46133937	15:98319310-98323713
9q22.31	9:90773659-90775181	9:73844298-73845820	9:87638489-87640007	1a:43252378-43252847	15:105703062-105704612
9q31.3	9:110273469-110275237	:93802632-93804400	9:107746343-107748108	1a:23924668-23926070	x
10p12.2	10:23660453-23667319	10:26418250-26424027	10:24274004-24282631	9:94568696-94576449	x
10q11.22	10_GL391668_random:1 -2595798	10:60441309-60443018	10:46196171-46197879	18:21298564-21300257	x

10q21.2	10:59416726-59423883	10:73291467-73297634	10:74741319-74747451	18:10202870-10207223	9:74099953-74116042
10q21.3	10:62431198-62432057	10:76370427-76371286	10:71638473-71639333	18:13285213-13286071	x
10q23.1	10:83921522-83926050	10:97559072-97563601	10:50387506-50393489	18:41641615-41647608	9:50613262-50618743
10q23.33	10:94267477-94273803	10:108095082-108103361	10_random38326859-38334121	3:39507006-39513297	x
10q24.1	10:95173385-95180892	10:109024442-109034593	10:94632970-94645543	3:38705576-38712611	9:92809544-92816736
11p15.4	11:9163630-9165442	11:9461344-9463157	11:60934646-60936458	x	8:56668618-56670407
11p14.3a	11:22149622-22154953	11:22663113-22667953	11:47628288-47630183	15:84260932-84262522	x
11p14.3b	11:25860690-25864367	11:26340528-26344195	11:43590825-43599036	15:87676434-87680167	14:46082361-46086411
11p14.2	11:26452534-26459761	11:26905367-26912602	11:43026432-43033620	15:88272823-88278334	14:45470380-45479191
11p12	11:38632668-38638317	11:39244620-39250273	11:30538323-30543825	15:100301380-100307140	14:33546556-33552218
11q14.1	11:75665777-75670940	11:74853102-74857792	11:73358167-73362396	15:59660949-59666209	14:76375851-76377787
11q14.2	11:84772239-84774880	11:84127443-84130732	11:82641610-82641633	11:82640564 -82643832	14:86620768-86622267
11q22.3	11:105837977-105839976	11:105659184-105661186	11:104606222-104608223	15:27666681-27668684	x
11q23.3	11:115950717-115951767	11:115955199-115956266	11:114968086-114969133	15:17136980-17138027	14:118969442-118969856
12p13.31a	12:7421458-7426840	12:7374554-7379948	12:7449851-7455274	23:27427620-27432991	11:7450259-7455372
12p13.31b	12:8986666-8993714	12:8995971-9002992	12:8796414-8803436	23: 26262780-26267054	11:8870627-8871734
12p11.1	12:51432590-51435497	12:34528554-34530759	x	x	x
12q12a	12:51007649-51010493	12:36151689-36153822	12:37751370-37754206	11:61847034-61849479	x
12q12b	12:50827568-50845105	12:36324031-36329077	12:37921282-37926328	11:62083673-62087011	x
12q12c	12:49060340-49066247	12:38091016-38096940	12:39733966-39739931	11:63869370-63875298	11:37186774-37192708
12q13.12	12:38404587-38415464	12:48875814-48886817	12:50601200-50612671	8:1173607-1174114	11:48189638-48190396
12q13.3	x	x	x	x	x
12q14.1	12:30483004-30487209 ?	12:57075987-57083864	12:58727824-58732584	11:53760680-53768957	x
12q21.31	12:84869627-84874320	12:83616034-83620718	12:85364687-85374072	10:20679266-20683975	11:82261770-82266494
12q23.3	12:105265980-105266770	12:104353656-104353940	12:106696108-106696898	10:78361081-78361871	11:106712091-106712892
12q24.31	12:124404096-124405024	12:123448427-123458820	12:125941849-125950836	10:97518000-97518457	11:125550634-125553778
12q24.33	12:132856209-132866534	12:131987153-131994803	12:134924778-134934488	10:105607947-105610978	11:133650206-133657152
13q13.3	13:36581610-36583829	13:18882604-18884816	13:36742483-36744698	9:2512263-2514472	17:16316350-16317501
13q21.1	13:54719736-54728073	13:37359656-37368114	13:55526951-55534627	5:82691956-82698382	x
13q21.31	13:64411775-64413117	13:47183274-47186257	13:65603074-65608526	x	17:44593353-44593958
13q21.33	13:69023873-69027576	13:51773898- 51777605	13:70266320-70269997	5:96593558-96594676	x
13q31.1	13:8248687-82491282	13:65326613-65330891	13:83980648-83985132	5:110263806-110268206	x
13q31.3	13:93221734-93223115	13:75987986-75989365	13: 95219848-95222182	5:120535132-120536511	17:73583665-73584655
14q11.2	14:21143479-21151503	14:3219728-3227348	14:21691360-21705202	22a:42033093-42033712	x
14q12	14:25112227-25113487	14:7271225-7272516	14:25651574-25652864	22a:38124291-38125548	x
14q21.2	14:43911175-43915597	14:26297666-26300718	14:45255134-45259343	1a:118612648-118616852	7:108104632-108108652
14q22.1	14:52312219-52313581	14:34117797-34119177	14:53872195-53873778	1a:85352335-85353689	x
14q23.1	14:57139354-57142656	14:38973657-38977456	14:58736786-58740805	1a:90217909-90221773	7:121144190-121151720
14q32.11	14:90761892-90763113	14:72895403-72896617	14:92495918-92497093	22a:14954639-14955838	x
15q21.3	15:52611980-52615068	15:34338315-34345813	15:52093893-52100022	6:45033272-45040813	7:32473716-32480926
15q22.32	15:64447073-64454472	15:46305937-46318140	15:64125918-64133109	6:86471361-86478877	7:45213172-45220099
15q26.1	15:89678346-89682398	15:72163693-72167740	15:89058017-89062011	6:111478511-111482540	7:71533136-71537388
17q12a	17:19878760-19885707	5:46782069-46785829	17:32134376-32138116	Un_GL397432_1:2248579-2249735	16:49334274-49337803
17q12b	17:21405311-21406586	5:48536222-48537496	17:30318122-30319393	x	x
17q21.33	17:48990226-48996979	5:33618168-33619680	17:42138735-42145322	14:913407-916411	16:36618536-36624505
17q22	17:53672733-53679754	17:37251275-37252937	17:37251271-37255227	14:5499089-5504411	16:41266770-41273748
18p11.31	18:11932303-11934533	18: 4702709-4713709	18:27764508-27776123	4:64871603-64872383	x

18p11.21	18:2869751-2870907	18:13896621-13897777	18:18384250-18385413	4:74091388-74092542	18:1065459-1066610
18q21.32	18:56809626-56812741	18:58854677-58858435	18:73462971-73466079	4:19418910-19420812	x
18q21.33	18:58584946-58585695	18:60660794-60661543	18:75297202-75297953	4:17623946-17624695	x
19p12a	19:20430843-20432775	19:20502302-20506171	19_random:5568523-4693457	x	x
19p12b	x	19:21998786-22004190	x	x	x
19p12c	19:22816285-22822589	19:22985546-22992205	19:22510578-22517155	10:52140975-52147928	x
19p12d	19:23958878-23959948	19:24176389-24177441	19:24090796-24091846	10:51086905-51087978	19:22963633-22964695
19q13.2a	19:45254473-45255205	19:37482886-37483591	19:41213813-41214527	17:86880260-86880974	19:46537460-46538166
19q13.2b	19:46167307-46169213	19:38376691-38376840	19:42109793-42111666	17:87714865-87715814	19:47519216-47520236
20p13	x	20:171220-174467	20:28225859-28229103	13:33687779-33691025	x
20q13.2	20:52490168-52494218	20:53454136-53458201	20:53286044-53287693	13:9174168-9178224	x
21q21.1	21:5208083-5215934	21:7083992-7092153	21:18606390-18615277	25:6004473-6016607	x
21q21.3	21:13034026-13041488	21:15153812-15159936	21:27348201-27348945	25:13435759-13442963	3:19971752-19973250
21q22.2	21:25817510-25824651	21:28294079-28301215	21:40844569-40852339	25:26213031-26219937	x
22q12.3	22:32628483-32628940	22:18172024-18172481	22:29109344-29109782	7b:16484326-16484772	x
Xp22.31	X:7534433-7539472	X:7472642-7477514	X:7397116-7402013	X:5561932-5567433	X:5203076-5205778
Xp11.3	X:46686230-46688943	X:46927369-46930082	X:46935626-46938572	X:45453723-45456398	X:44969444-44972003
Xp11.21	X:58147265-58153612	X:58316397-58323520	X:57907519-57914952	X:35724314-35731339	X:56463748-56466180
Xq12	X:65895262-65897531	X:63496539-63498812	X:63676878-63679215	X:59687935-59690037	X:65509808-65511881
Xq13.3	X:75719292-75724090	X:73047143-73053922	X:73343759-73348774	X:68207999-68214772	x
Xq21.1a	X:78189675-78191076	X:75442112-75446023	x	X:69944317-69946389	x
Xq21.1b	x	X:77046330-77053487	X:77756962-77764114	x	x
Xq22.3a	x	X:103289509-103292744	X:104958367-104961600	X:93503162-93506222	X:105123999-105124946
Xq22.3b	X:107494230-107496958	X:104359006-104361742	X:106007418-106012119	X:94498278-94500999	X:106203542-106206248
Xq23	X:117209721-117211140	X:113981816-113985099	X:115682386-115684958	X:103792112-103794673	x
Xq26.2	X:132530228-132534002	X:129455032-129458532	X:131362733-131363956	X:118464319-118465912	X:131248979-131252737
Xq27.1	X:140805217-140808406	X:137749632-137752786	X:139728041-139731236	X:126598487-126601688	X:139347358-139350324
Yp11.2	Y:6464881-6471520	x	x	x	x
Yq11.222	Y:18689301-18696863	x	x	x	x

For each regarded (H)ERV-W locus, chromosomal bands (proviral loci indicated in bold, loci lacking both LTRs in italics) are given for the regarded locus in the in the human GRCh37/hg19 reference genome sequence (for precise start and end positions, see [121]. Chromosome coordinates of orthologous ERV-W loci are given for the other regarded non-human Catarrhini primate reference genome sequences. Apparent absence of a (H)ERV-W sequence in the orthologous genome position of other primate genome sequences is indicated by "x". Two HERV-W loci on the human chromosome Y were excluded from the analysis (see text).

Table 14. *Number of orthologous of HERV-W loci in the analyzed Catarrhini primate genome sequences*

	Chimp	**Gorilla**	**Orangutan**	**Gibbon**	**Rhesus**
HERV-W loci corresponding to human 211[a] HERV-W elements	205	207	205	190	131

[a] no comparative information available for the two HERV-W loci on Y chromosome

In addition to this first wave of ERV-W locus formations, a total of 80 HERV-W loci was lacking an orthologous locus in Rhesus, but had orthologs only in subsequent *Hominoidea* species (<30 MYa), suggesting a continuous formation of novel HERV-W loci (66). Differently, relatively few insertions, specifically 14, likely occurred later on, between 20 and 17 MYa, based on the comparative genomics data provided by UCSC Genome Browser (Figure 18).

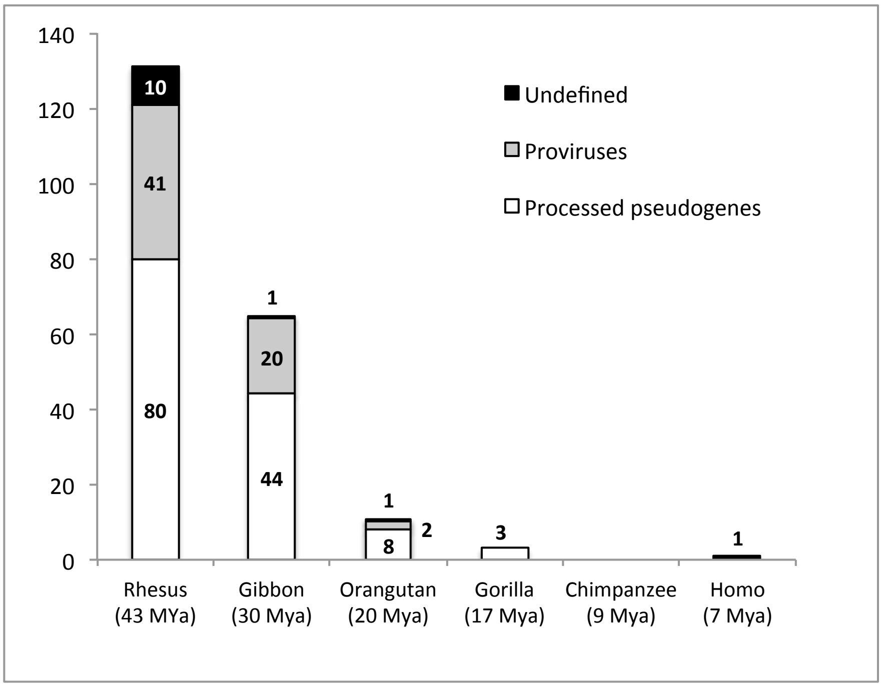


Figure 18. *Initial formation of 211 HERV-W loci based on presence of respective orthologous loci in Catarrhini primate reference genome sequences.*
The number of orthologs of HERV-W loci initially formed in a particular primate species is given for each species. HERV-W loci being either proviruses, or processed pseudogene generated by L1-machinery, or undefined elements (see text). For instance, 20 HERV-W loci in the human genome were initially formed in the common ancestor of human and Gibbon. Eight HERV-W loci representing processed pseudogenes, were formed in the common ancestor of human and Orangutan. Note that the majority of HERV-W loci was initially formed in the common ancestor of human and Rhesus and is thus common to all Catarrhini genomes, while a minority of HERV-W locus formations occurred later in Catarrhini primate evolution. Approximate time periods of last common ancestors of Catarrhini primate lineages are given in millions of years (Mya) below species names.

Overall, (H)ERV-W locus comparisons in primate genome sequences indicated that the ERV-W group formed new loci throughout an extended period of time during evolution, due to both novel proviral acquisitions (n=63) and, notably at even higher numbers, to L1-mediated processed pseudogene formations (n=133). In particular, >90% of ERV-W orthologous loci formation occurred in Rhesus (n= 131) and Gibbon (n= 65), approximately between 43 and 20 MYa, showing an approximately 2:1 ratio of L1-mediated processed pseudogene formation relative to provirus formations. These data indicate that formation of ERV-W loci being processed pseudogenes occurred at considerable levels and in parallel to "true" proviruses formation. This further suggests also that ERV-W transcripts serving as templates for ERV-W processed pseudogenes formation by L1-machinery must have been present in the germ line during that time period. As for Orangutan, our analysis indicated a pronounced decline of ERV-W locus formations, yet 8 ERV-W loci being processed pseudogenes versus 2 ERV-W loci being "true" proviruses shows that L1-mediated formation of ERV-W loci was predominant in the evolutionary lineage leading to Orangutan. Similarly, all the 3 ERV-W loci formed along the evolutionary lineage leading to Gorilla were ERV-W processed pseudogenes, further suggesting that the L1-mediated formation of ERV-W loci occurred for an extended period of time when compared to "true" provirus formations. Of further interest, no new formations of ERV-W loci were observed in Chimpanzee; while a HERV-W locus in human chromosome 12q13.3 appeared to be human-specific, with a gap in the corresponding genome regions of all non-human *Catarrhini* primates, thus initially suggesting that an HERV-W insertion has occurred <7 MYa, i.e. the estimated separation time between Homo and Chimpanzee [221, 222]). However, the human-specificity of this HERV-W sequence is uncertain given that the actual HERV-W sequence in chromosome 12q13.3 is of highly mutated structure and is lacking both LTRs also making sequence divergence-based age estimates very unreliable [21].

4.3 Analysis of ERV-W sequences identified by sequence similarity searches in non-human *Catarrhini* primates identifies species-specific insertions

The comparative analysis of the human orthologous ERV-W sequences found in *Catarrhini* primates revealed an extended period of ERV-W locus formations throughout primates evolution, with evidently 80 novel insertions since the evolutionary separation of Gibbon and Homo lineages. Such an extended time period of ERV-W activity could likely also has resulted in species-specific ERV-W insertions outside of the human evolutionary lineage. Hence, to identify potential species-specific ERV-W insertions lacking an orthologous locus in humans, we performed BLAT searches of *Catarrhini* primate genome sequences provided by UCSC Genome Browser using the HERV-W group LTR17-HERV17-LTR17 RepBase sequence as a query.

It is worth to note that the HERV17 BLAT search based approach identified, overall, a lower number of ERV-W loci in each non-human *Catarrhini* primate as compared to the number of ERV-W loci obtained from the above comparative analysis of orthologous of human ERV-W integrations. In particular, except for Rhesus, as compared to the total number of orthologous loci identified by comparative analysis of homologous genome regions, the BLAT-based strategy identified on average only the 79% of those ERV-W loci, suggesting a proportion of ERV-W elements not effectively detected by BLAT searches strategy (Table 15).

Table 15. *Number and human correspondence of HERV17 Blat-retrieved ERV-W sequences in the analyzed Catarrhini primates genomes*

	Chimpanzee	**Gorilla**	**Orangutan**	**Gibbon**	**Rhesus**
1) ERV-W loci with HERV-W orthologous in human genome	138 (67%)	132 (64%)	122 (60%)	111(58%)	69 (53%)
2) ERV-W loci corresponding to human solitary LTRs (n=19)	1 (17)	1 (17)	7 (10)	10* (8)	14* (0)
3) ERV-W loci present in human as non-canonical HERV-W (like)	29	27	24	21	20
4) ERV-W loci lacking an orthologous in human	3 (3)	5 (4)	8 (6)	4 (2)	68 (66)
TOTAL	**171**	**165**	**160**	**145**	**168**

1) Number of ERV-W elements in the various primate genome sequences with an orthologous locus among the 211 HERV-W loci in the human reference genome sequence. Respective percentages, relative to 211 HERV-W loci, are given in parenthesis. Two HERV-W loci on human chromosome Y were excluded from the analysis (see text).
2) Numbers of ERV-W proviral elements in the various primate reference genome sequences with a solitary HERV-W LTR (LTR17) at the orthologous human reference genome sequence position. Numbers in parenthesis indicate the amount of proviral insertions acquired in evolutionarily older primates species that corresponds to a solitary LTR also in the non-human primates analyzed. *Species of first acquisition of the proviral insertions leading to the formation of solitary LTRs in subsequent primates species: Gibbon (5) and Rhesus (14).
3) Numbers of ERV-W elements in the various primate reference genome sequences with an orthologous in the human reference genome sequence, yet with the ERV sequence being less similar to HERV-W. Those sequences were not identified as HERV-W elements in a previous analysis (Grandi et al. 2016)
4) ERV-W loci absent in the orthologous genome positions in the human reference genome sequence. Numbers in parenthesis indicate the proportion of species-specific insertions, that is, ERV loci not present in any of the other examined primate genome sequences.

We hence decided to further investigate those different outcomes, localizing the human orthologous genome regions corresponding to each ERV-W locus identified by BLAT. Only 53-67% of the ERV-W orthologs identified by the above comparative analysis of homologous genome regions (Table 14) were effectively identified by BLAT search (Table 15, row 1)) (Table 1). The remaining BLAT-identified ERV-W loci (without orthologous HERV-W loci) could be explained by three corresponding conditions in the human GRCh37/hg19 assembly: i) presence of a solitary LTR annotated as LTR17 (Table 15, row 2)); ii) presence of HERV-W-like elements with somewhat lesser identity to HERV17 (on average, ~63% pairwise identity) (Table 15, row 3)); iii) complete absence of HERV-W or HERV-W-like sequences (Table 15, row 4)). Each of those three conditions was analyzed separately, and results are described in the following.

4.3.1 ERV-W BLAT-identified sequences being solitary LTRs in humans

In a minority of cases, a solitary LTR annotated as LTR17 was found at the orthologous position in the human genome sequence (Table 15, row 2)), suggesting a previous event of homologous recombination between the 5′ and 3′ LTRs, that eliminated the internal proviral portion [5]. The genomic positions of those 19 human solitary LTRs and their corresponding positions in non-human Catarrhini primates are reported (Table 16).

Table 16. *ERV-W loci in non-human Catarrhini primates genomes with a solitary HERV-W LTR at the orthologous genome position*

Homo solo LTR	Chimpanzee	Gorilla	Orangutan	Gibbon	Rhesus
1:192855545-192856320	1:171523796-171524572	1:172632606-172633382	1:57461588-57462385	**9:23097469-23105282**	x
2:157501270-157502058	2B:160868497-160869285	2B:160868497-160869285	2B:44160134-44160922	**17:21605862-21613345**	x
4:190972713-190973051	x	4:201094988-201095326	x	*7b:115087284-115088896*	x
5:89713141-89713870	5:25005173-25005902	5:73246322-73247047	**5:90947542-90953989**	**2:104500510-104506980**	*6:87070391-87075542*
5:114427938-114428650	5:115582850-115583562	5:98189870-98190631	**5:116094487-116101834**	**2:129150454-129157611**	**6:112385135-112392351**
5:114449574-114449785	5:115604449-115604660	5:98211617-98211828	**5:116137410-116144419**	**2:129188751-129195820**	**x**
7:27315710-27316490	7:25865229-25866009	7:27315710-27316490	*7:57158928-57169840*	x	**3:99520876-99530859**
7:80811340-80812054	7:81677947-81678661	x	x	*11:12139249-12145059*	**3:137158373-137165344**
7:119018647-119019408	7:120883617-120884378	7:117327055-117327818	7:116127440-116128201	13:70613353-70614227	**3:158670031-158675988**
7:125095615-125096399	7:126995440-126996220	7:123528656-123529436	7:122379167-122379969	13:76714580-76715328	*3:164521606-164528889*
13:78633216-78633978	13:78034364-78035126	13:60830048-60830810	13:79495266-79496031	**5:105516008-105523568**	**17:58566647-58572935**
13:100186663-100186733	13:99772916-99772986	13:82523858-82523928	13:101878590-101878660	5:127460778-12746084	**17:80094843-8010162**
14:53574468-53575226	14:52063696-52064454	14:33864565-33865323	**14:53613221-53622798**	1a:85103619-85104372	**7:116144166-116150056**
15:45839251-45839993	15:42696744-42697486	15:24303530-24304272	15:42001410-42002138	6:55113552-55114255	**7:22728272-22734350**
15:78531314-78532373	15:75828303-75829356	15:57975329-57976369	15:75816660-75817723	6:97797145-97798212	*7:56712211- 56721682*
17:74608050-74608803	17:75582200-75582806	5:6606848-6607601	*17:66684744-66689032*	**14:92032349-92038770**	*16:74196820-74198756*
18:47975521-47976284	**18:46380751-46388713**	**18:48193528-48195657**	**18:62832726-62846974**	**4:29802680-29810316**	x
18:63937324-63938112	18:62389950-62390723	18:64568234-64569016	18:79266131-79266915	4:13879305-13879701	**18:60615570-60624170**
X:49480900-49481607	X:49807850-49808558	X:50383089-50383796	X:50249996-50250703	X:42243207-42243895	**X:48179044-48185277**

Chromosome coordinates of human solitary HERV-W LTRs in the GRCh37/hg19 reference genome sequence are given in the first column. Chromosome coordinates of orthologous genome regions are given for each primate's reference genome sequence. Coordinates given in bold indicate a more complete ERV-W locus consisting of LTR-internal portions–LTR. Coordinates given in italics indicate ERV-W loci lacking one LTR. Apparent absence of an ERV-W sequence in the orthologous genome position of other primate genome sequences is indicated by "x".

Each human solitary LTR derived from an ERV-W proviral integration that had occurred either in Rhesus (14) or Gibbon (5), in line with the ERV-W group's main period of germ line colonization. None of the solitary or proviral LTRs or respective loci showed signatures of processed pseudogenes, that likely would have prohibited homologous recombination due to relatively short homologous sequences within 5' and 3' LTR portions present.

4.3.2 ERV-W BLAT-identified sequences corresponding to HERV-W-like elements with lesser identity to HERV17.

The now identified lower scoring HERV-W-like elements (Table 15, row 3) and Table 17) had not been identified as HERV-W loci by BLAT searches during the characterization of the group in the human genome (Chapter 3 and ref. [21]). The closer inspection of RepeatMasker annotations revealed that some of those loci were annotated in the GRCh/hg19 assembly as composed of stretches of other γHERVs, such as HERV9, HERV30 and HERVIP10FH regarding internal portions, and LTR12F regarding LTR portions, while they were annotated as HERV17 in the non-human primates genome sequences. Also, some of the loci were previously identified as non-canonical HERV9 elements that are closely related to HERV-W [8].

Interestingly, approximately two-thirds of the concerned HERV-W-like elements are present at respective orthologous positions in species ranging from Rhesus to human. Thus, those loci likely were formed during the main period of the HERV-W group's colonization of primate genomes. Some of the elements apparently lacking in some intermediary primate species potentially suggests species-specific deletion or rearrangement events. The remaining approximately one-third of HERV-W-like elements presumably entered primate genomes only in the evolutionary separated lineages leading to Gibbon (3), or Orangutan (2), or Gorilla (2), while no novel elements were observed for Chimpanzee, the latter as already observed for HERV-W orthologous loci.

***Table 17**. ERV-W loci in non-human Catarrhini primates with HERV-W-like elements with lesser similarities to HERV-W*

HUMAN	St[a]	COMPOSITION [b]	Chimpanzee	Gorilla	Orangutan	Gibbon	Rhesus
1:47412274-47413748	+	HERV17	1:47319233-47320712	1:48563504-48564989	1:182883772-182885110	x	x
1:47595486-47596960	-	HERV17	1:47507094-47508583	1:48762353-48763600	x	x	x
2:180922439-180924977	+	HERV17	2B:184557730-184560322	2B:68131599-68134157	2b:70378703-70381263	22a:70206288-70208838	12:43412226-43415355
3:88935018-88936153	-	LTR17, HERV17	3:90518902-90520082	3:91061210-91062323	3:56060024-56061141	21:30570196-30571307	x
3:109693465-109693986	-	LTR17, HERV17	3:113292158-113292683	3:109243100-109243621	x	x	2:30136011-30136601
5:74600280-74602253	+	LTR17, HERV17	5:40246349-40248322	17:22481773-22483748	5:75467675-75469638	x	6:71906965-71913523
5:34062285-34064918	-	HERV17, HERVIP10F, LTR17	5:81115686-81118310	17:59317826-59318,681	5:35260,596-35261453	6:33094516-33097177	6:34661967-34663609
5:18742581-18746140	-	HERV17	5:96642151-96645792	Assembly gap	5:19527081-19530715	6:17903593-17906353	6:18906784-18910540
6:86683575-86685615	-	LTR12F, HERV17, HERV30	6:86562949-86564933	6:85991558-85993594	6:86876026-86878099	3:72763651-72765735	4:84026123-84028150
6:115319550-115323003	+	HERV17	6:116101700-116106563	6:115240966-115244473	6:117282305-117286633	3:102243303-102246147	4:151174946-151179630
6:166759559-166760142	-	HERV17	6:168227391-168228402	6:167692411-167692997	6:170015596-170016132° 6:170026363-170026946°	3:153964493-153964928	4:165517503-165519944
7:9057073-9059465	+	HERV17	7:7581216-7583603	7:9096693-9098735	7:75998958-76001318	11:31271584-31273908	3:117908936-117911567
8:124422873-124424005	+	LTR17, HERV17	8:122082425-122083557	8:123166377-123167482	8:131091812-131093080	16:78319169-78320301	x
8:129390098-129394380	-	LTR12F, HERV17	8:127056725-127061561	8:128190281-128194304	8:136150893-136155139	16:83391362-83392698	8:133381887-133385160
9:29776755-29779436	-	HERV17	9:30136632-30139344	9:30316499-30319231	9:31880122-31883086	1a:71137929-71140557	15:48196069-48198624
10:99049848-99051021	+	LTR17, HERV17	10:96770118-96771283	10:110637802-110638975	x	3:37096811-37097976	x
11:18840552-18842462	-	HERV17	11:18604731-18606655	11:19066602-19068576	11:51298765-51303323	x	x
11:114217924-114220414	-	HERV17	11:112256075-112258622	11:112196590-112199115	11:111167151-111169570	15:20961640-20964130	14:115346686-115349361
12:14369478-14374042	+	LTR12F, HERV17	12:14489812-14495927	12:14360759-14365091	12:14657239-14661331	23:21219200-21224126	11:14483836-14488191
12:18220942-18223574	+	HERV17	12:18382860-18385470	12:18293700-18296333	12:18783066-18785637	23:17456743-17459320	11:18363379-18365928
14:38573435-38576501	+	HERV17	14:37075265-37078756	14:19443428-19446495	14:37874964-37878284	1a:112706086-112710236	7:100951196-100955957
15:51649330-51653005	+	LTR12F, HERV17	15:48582834-48586513	15:30245167-30247494	15:47955747-47959248	6:49161294-49164720	7:28486674-28491756
16:18124951-18125497	-	LTR17, HERV17	16:17959662-17960501	16:18736143-18736687	x	x	x
16:20656274-20658178	+	HERV17	16:20490450-20492229	16:21236747-21238710	16:20007845-20010152	2:89380068-89381930	20:19578770-19591430
18:1729227-1730327	-	LTR17, HERV17	18:14991994-14993087	18:1739593-1740701	18:30978591-30979682	4:61689388-61690488	18:13661678-13662432
19:24350451-24353078	-	HERV17	x	19:24402896-24405539	19:24304705-24307297	10:50846881-50849462	19:23526703-23527188
X:105794494-105796487	+	LTR17, HERV17	X:107003370-107004720	X:103861433-103863410	X:105490168-105492161	X:93999146-94000864	X:105707949-105709670
Y:19244538-19249067	-	LTR17, HERV17, HERVH, HERVIP10FH	Y:19256482-19260910	-	-	-	-
Y:19246861-19247702*	+	HERVIP10FH	Y:19257701-19258635	-	-	-	-
Y:16452374-16456884	-	LTR12F, LTR17, HERV17, HERV30	Y:21952136-21956101	-	-	-	-

Chromosome coordinates of human HERV-W LTRs in GRCh37/hg19 sequence are given in the first column. [a]Strandedness and [b]Repeatmasker annotation of respective HERV regions are given in the following columns.

Chromosome coordinates of orthologous genome regions are given for each primate's reference genome sequence. Apparent absence of an ERV-W sequence in the orthologous genome position of other primate genome sequences is indicated by "x". Two HERV-W loci on the human chromosome Y were excluded from the analysis (see text).

° An orthologous sequence in Orangutan chromosome 6:166759559-166760142) is present in two identical copies.

* An orthologous ERV-W sequence in chromosome Y:19246861-19247702 is subsequent antisense integration into the ERV-W in chromosome Y:19244538-19249067.

4.3.3 ERV-W BLAT-identified sequences lacking an orthologous in human.

A number of ERV-W sequences identified by BLAT searches in non-human *Catarrhini* primate genomes lacked orthologous loci in the human genome (Table 15, row 4) and Table 18). In theory, such ERV-W elements may have formed species or lineage-specifically, and thus they could also provide information on the ERV-W group time period(s) of activity. For example, ERV-W loci that are present exclusively in Rhesus could have formed after separation of the Rhesus evolutionary lineage from the other primates lineages, about 30 MYa, while sequences present exclusively in Gibbon, Gorilla or Chimpanzee could indicate an integration event in the primate lineage after the evolutionary separation of respective lineages, about 20, 17 and 9 MYa, respectively [221, 222] (Figure 17). Interestingly, as also reported in Table 18, the great majority (81/88) of BLAT-identified ERV-W sequences that lack orthologs in human are actually species-specific insertions, further suggesting an extended period of ERV-W germ line colonization in primates. In particular, for the Rhesus genome, 77% of ERV-W insertions appeared to be absent in humans, with still 66/68 species-specific elements when compared to primate species more closely related to Rhesus. This further indicates that the main period of ERV-W activity ranges from 43 MYa to less than 20 MYa, with a greater number of Rhesus specific ERV-W element formations that occurred after separation of the Rhesus evolutionary lineage. The other non-human *Catarrhini* primate genomes likewise showed some evidence for ERV-W insertions lacking a human orthologous (Table 15, row 4) and Table 18): Gibbon: 4 insertions (2 species-specific), Orangutan: 8 (6 species-specific), Gorilla: 5 (4 species-specific), and Chimp 3 species-specific insertions.

Noteworthy, the species-specific insertions found in Rhesus and Gorilla included a number of new proviral sequences (15 and 1, respectively) (Table 18). This suggests that the ERV-W species-specific diffusion in Rhesus and Gorilla genomes has been in part due to the acquisition of new proviruses, by either intracellular provirus formations, or potentially re-infections, and likely hinting at sporadic formation of novel elements during the last 10-5 MY. Moreover, species-specific formations of processed pseudogenes of ERV-W strongly suggest that L1-mediated mobilization of ERV-W sequences has been ongoing for considerable time periods also in primates outside of the human evolutionary lineage. In fact, species-specific formations of processed pseudogenes were observed in Rhesus (24), Orangutan (3), Gorilla (1) and Chimpanzee (1), indicating novel retrotransposition of ERV-W elements by L1-machinery in various primate lineages, between about 43 and 5 MYa.

Table 18. *ERV-W loci in non-human Catarrhini primates lacking an orthologous in human genome sequence*

ERV-W loci	**Strand**	**Length**	**Type***	**Orthologous loci**
Chimpanzee				
1:28472456-28472713	-	258	UN	x
12:54255388-54258420	-	3033	PG	x
19:10251447-10252008	-	562	UN	x
Gorilla				
6:137848377-137848999	-	623	UN	x
9:99983329-99984432	+	1104	?	x
12:35692937-35695882	-	2946	PG	x
12:38035777-38041701[a]	+	5925	PV	x
17:59317825-59318777	+	953	PG	Orangutan 5:35260596-35261453
Orangutan				
5:29323056-29325960	+	2905	PG	Gibbon 6:27609950-27613887 Rhesus 6:28927062-28930510
6:116136923-116139893	+	2971	PG	x
6:169988045-169988778	+	734	UN	x
10:6196575-6198307	+	1733	UN	x
12:55188568-55189227	+	660	UN	x
14:26833423-26836116	+	2694	PG	Gibbon 22a:36931444-36934101[b] Rhesus 7:90073203-90077221 [b]
14:26842121-26842940	+	820	PG	x
19:40880584-40883048	-	2465	PG	x
Gibbon				
6:27609950-27613887	+	3418	PG	Orangutan 5:29323056-29325960 Rhesus 6:28927062-28930510
11:57319215-57321839	-	2625	UN	x
20:58589539-58590163	+	625	UN	x
22a:36931444-36934101	-	2658	PG	Orangutan 14:26833423-26836116 Rhesus 7:90073203-90077221 [b]
Rhesus				
1:1038263-1047401	-	10491	PV	x
1:49193091-49198666	-	5576	PG	x
1:51518080-51521456	-	3377	UN	x
1:51551811-51557699 [c]	-	5889	UN	x
1:80552272-80559161	+	6990	PV	x
1:104855590-104856392	-	803	UN	x
1:136470672-136473027	-	2356	PG	x
1:200071013-200079914	-	8902	PV	x

2:78537-80690	+	2154	PV	x
2:23026824-23028023	-	1200	UN	x
2:45734713-45736546	-	1834	PG	x
2:56018048-56021863	-	3816	PV	x
2:79150722-79152714	-	1993	UN	x
2:81235387-81239098	-	3712	UN	x
2:89484049-89484553	+	505	UN	x
2:97124678-97133786	+	9109	PG	x
2:150753213-150760402	-	7190	PG	x
2:187806335-187807033	+	699	UN	x
3:79142677-79148871	-	6195	PG	x
3:113453986-113459310	+	5325	UN	x
3:134506376-134514022	-	7647	PG	x
3:147768123-147772294	-	4172	PG	x
4:4004556-4011519	-	7036	PG	x
4:128350042-128350560	+	519	UN	x
4:129174758-129176226	+	1469	UN	x
5:15752920-15759463	+	6544	PV	x
5:101699598-101700252	+	656	UN	x
5:138607909-138608458	+	550	UN	x
6:27115997-27119842[d]	+	3847	UN	x
6:28927062-28930510	+	3450	PG	Orangutan 5:29323056-2932596 Gibbon 6:27609950-27613887
6:34074975-34080489	-	5516	PV	x
6:58804908-58808717	+	3810	PG	x
6:107032014-107035698	-	3685	UN	x
7:18098546-18100122	+	1577	PG	x
7:41800821-41803587	+	2767	UN	x
7:68361812-68362489	-	678	UN	x
7:90073202-90077221	+	4020	UN	Orangutan 14:26833423-26836116 Gibbon 22a:36931444-36934101 [b]
7:134568182-134575148	-	6967	PG	x
7:145163419-145164942	-	1524	PG	x
7:145174506-145180394	-	5890	PG	x
8:28503905-28505303	+	1400	UN	x
8:111644122-111647744	-	3623	PV	x
9:50627278-50628617	+	1340	PV	x
9:61574841-61579967[d]	-	5127	UN	x
9:76546813-76548675	+	1863	PG	x
9:84724356-84727832[d]	-	3468	UN	x
9:124054089-124056102	-	2014	UN	x
10:46065109-46066965	-	1857	PG	x
10:62702712-62705095	-	2384	PG	x
11:52843688-52850515	+	6829	UN	x
11:56735387-56742250	+	6864	PV	x
11:78399468-78405509	-	6042	PG	x
12:78740985-78746777	+	5793	PV	x

14:48154691-48162700	+	8010	PG	x
14:64349790-64351688	-	1899	UN	x
14:94472347-94480147	-	7801	PV	x
15:45730394-45737226	-	6833	PG	x
15:105345557-105352812	-	7256	PG	x
16:32790950-32796478	+	4718	PG	x
17:35362274-35367596	-	5323	PV	x
17:43147477-43151327[d]	-	3851	UN	x
17:70149164-70152579	-	3416	UN	x
X:75422858-75434413	-	11557	UN	x
X:75438492-75439743	-	1252	UN	x
X:85447809-85449859	-	2051	PG	x
X:85457714-85458989	-	1276	UN	x
X:146260510-146265185	-	4676	PV	x
X:146271192-146277072	-	5881	PV	x

Chromosome coordinates of ERV-W loci are given as chromosome:start-end positions with respect to each primate reference genome sequence.

[*] ERV-W locus types are given as PV = provirus, PG = processed pseudogene, UN = undefined (see text).

[a] A duplication of 12:38091016-38096940 (corresponds to 12q12c in homo).

[b] Gibbon and Rhesus sequences showed a L1 insertion not present in Orango and partially present as a 55 nts portion in Gorilla and Chimpanzee.

[c] related to MER and PABL_A based on further analysis.

[d] related to HERV-H based on further analysis.

4.4 Identification of sequences closely related to HERV-W in *Platyrrhini* (New World Monkeys)

As mentioned above, a BLAT search with an HERV17 query sequence in Marmoset (*Callithrix jacchus*) and Squirrel Monkey (*Saimiri boliviensis*) NWM genome sequences did not identify true ERV-W insertions in those primates (data not shown), confirming that the spread of the ERV-W group has been limited to *Catarrhini*.

Interestingly, however, the BLAT searches identified a group of apparently highly related multi-copy sequences in *Platyrrhini*, annotated in the respective Genome Browser reference sequence as ERV1-1_CJa-I for the internal portion and as ERV1-1_CJa-LTR for the 5′ and 3′ LTRs. For brevity, those sequences will be referred to as ERV1-1 in the following text.

Sequence similarities of HERV-W and the identified ERV1-1 elements were further examined at the nucleotide level (Figure 19). Comparison of a typical HERV-W proviral sequence, consisting of LTR17-HERV17-LTR17, with the ERV1-1 sequence, as included in RepBase, revealed an overall 73% sequence identity between internal *gag*, *pro*, *pol* and *env* gene portions (approximately nucleotides 2700 to 7750 in the HERV-W proviral sequence). A portion of the HERV-W *env* gene, spanning approximately from nucleotide 7750 to 8570 in the HERV-W reference sequence, seems to be absent in the ERV 1-1 reference (Figure 19A).

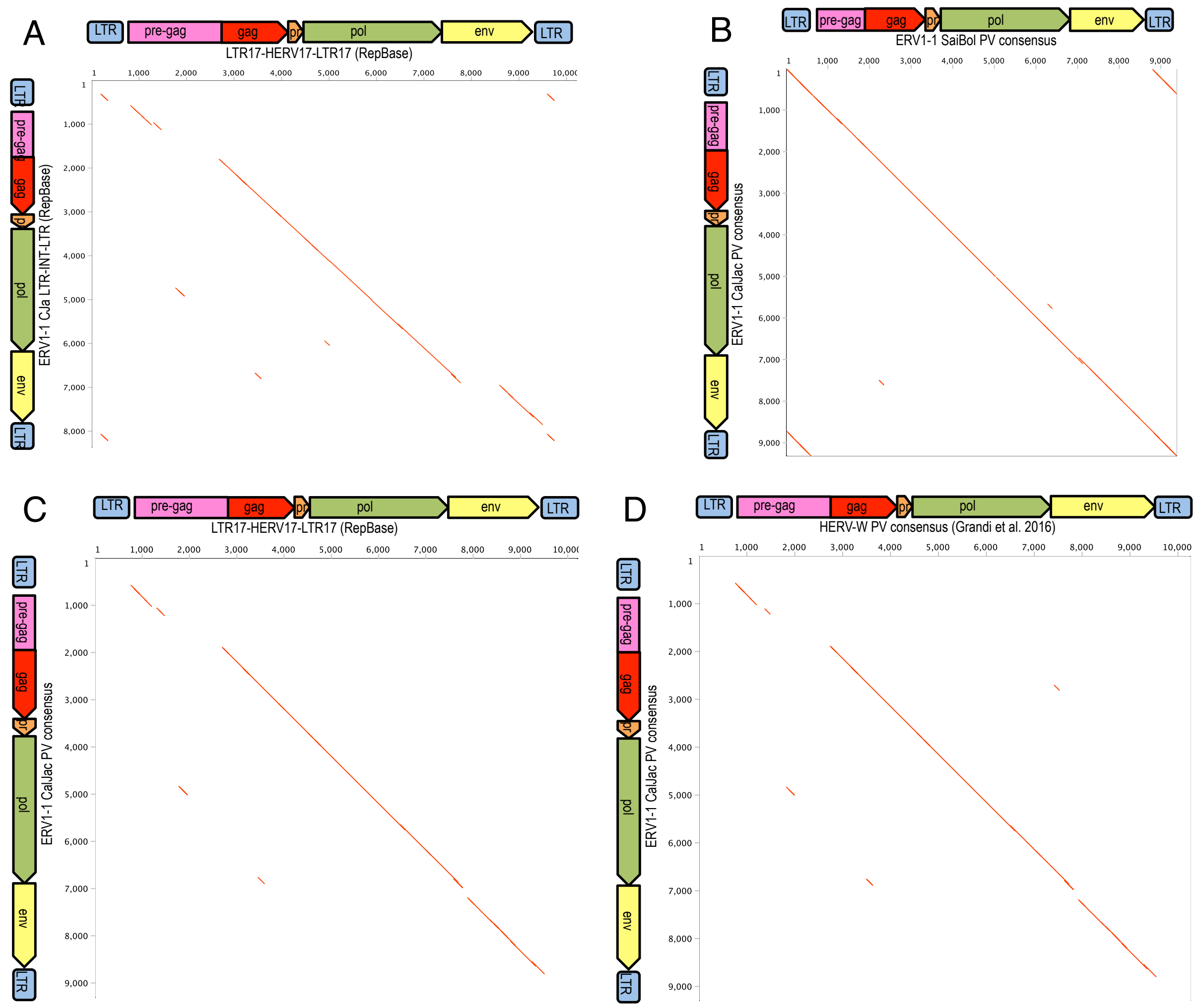


Figure 19. *Pairwise nucleotide sequence comparisons depicting similarities between HERV-W and other ERV-W sequences.*

Reference sequences and consensus sequences were compared with each other as follows: A) Callithrix jaccus ERV1-1 RepBase sequence and HERV-W RepBase sequence; B) Callithrix jaccus and Saimiri boliviensis ERV1-1 proviral consensus sequences as generated in this study; C) Callithrix jaccus ERV1-1 proviral consensus as generated in this study and HERV-W RepBase reference sequence; D) Callithrix jaccus ERV1-1 proviral consensus sequence as generated in this study and a HERV-W proviral consensus as generated in Chapter 3 and reported recently [21]. Sequence similarities in dot-plot comparisons are highlighted for sequence regions with at least 50% similarity along a 100 nt sequence window. Proviral gene and LTR regions of ERV1-1 and HERV-W sequences are depicted.

We further investigated whether that difference within the *env* gene is typical for ERV1-1. We retrieved sequences of reasonably complete ERV1-1 proviral sequences (based on chromosome coordinates obtained from BLAT searches plus 5kb of upstream and downstream flanking sequence each) to compile an as comprehensive as possible collection of ERV1-1 proviral sequences. The collected ERV1-1 proviruses were analyzed for presence of 5' and 3' LTRs. Complete ERV1-1 proviral sequences from Marmoset (59 elements) and

Squirrel Monkey (71 elements) genome assemblies were used to generate two species-specific multiple sequence alignments and, subsequently, two majority rule-based consensus sequences, named ERV1-1_CalJac_PVconsensus and ERV1-1_SaiBol_PVconsensus, respectively. Those consensus sequences were subjected to by dot-plot comparisons and pairwise alignment to assess differences between the ERV1-1 group in the two NWM species (Figure 19B). Since the two consensus sequences showed only minor differences (98% overall identity), the ERV1-1_CalJac proviral consensus sequence was chosen as representative for both species ERV1-1 elements for subsequent analysis. When comparing the ERV1-1_CalJac proviral consensus with the LTR17-HERV17-LTR17 (Figure 19C) and with the HERV-W consensus built from the human proviral dataset [21] (Figure 19D), the above mentioned *env* portion not present in the ERV1-1 RepBase reference sequence was found to be due to a large deletion within that *env* gene in the majority of ERV1-1 sequences, as also observed for about 80% of HERV-W elements in the human genome (Chapter 3 and ref. [21]). Addition of this missing *env* portion in the ERV1-1 proviral consensus sequence thus confirmed high sequence identity with HERV-W also along the entire (full-length) *env* gene. Interestingly, sequence comparisons between ERV1-1 and HERV-W further showed that ERV1-1 sequences also harbor a so called *"pre-gag"* region that is located between the 5'LTR and the *gag* gene, as previously reported for HERV-W sequences (approximately nucleotide 800 to 2700 in LTR17-HERV17-LTR17) (Chapter 3 and ref. [21]). Of further note, ERV1-1 LTRs did not show pronounced similarity (overall 34%) with either the LTR17 RepBase sequence or the HERV-W LTR PV consensus, the latter as generated in Grandi et al. [21]. In further accord with that finding, BLAT searches did not identify sequences structurally resembling LTR17 in Marmoset or Squirrel Monkey genome sequences.

4.5 Presence of ERV-W related elements in other NWM families

To the best of our knowledge, unlike Marmoset and Squirrel Monkey, no genome sequence assemblies are available for species of the other two *Platyrrhini* families, *Atelidae* and *Pitheciidae*. We therefore probed the NCBI Trace Archive sequence database by performing BLAST searches restricted to sequences from Spider Monkey (*Ateles geoffroyi*, *Atelidae*) and Red-bellied Titi (*Callicebus moloch*, *Pitheciidae*), using LTR17-HERV17-LTR17 and the ERV1-1_CalJac generated proviral consensus sequence as queries. Both query sequences identified basically the same sequence entries, confirming the presence of ERV elements highly related to ERV-W when regarding proviral gene regions. The BLAST search with the ERV1-1 proviral consensus sequence produced overall higher identity scores along the entire proviral sequence (data not shown). This analysis therefore demonstrated the presence of ERV-W related ERV1-1 elements also in the other two NWM families, *Atelidae* and *Pitheciidae*, further detailing the spread of ERV1-1 in the parvorder *Platyrrhini*.

4.6 Absence of elements closely related to ERV-W in Prosimians

To complete our search for sequences closely related to ERV-W within the non-human primates, we performed BLAT searches in *Tarsiiformes* and *Prosimians* genome sequence assemblies available at the UCSC Genome Browser, i.e. Tarsier (*Tarsius syrichta*) and both Bushbaby (*Otolemur garnettii*) and Mouse Lemur (*Microcebus murinus*), respectively, yet retrieving only short sequence matches with insignificant scores (data not shown).

This provides further evidence that the spread of HERV-W related elements took place after the evolutionary separation of the *Prosimians* and *Simiiformes* lineages, occurred approximately 60 MYa [221, 222].

4.7 Analysis of retroviral genes and derived puteins corroborate close relationship of ERV1-1 with the ERV-W group

To further characterize sequence relationships between ERV1-1 and the ERV-W group, we performed pairwise comparisons of the ERV1-1 consensus nucleotide sequence with the consensus of HERV-W, HERV9 and HERV30, revealing the highest overall similarities with ERV-W, yet still considerable similarity of ERV1-1 with HERV9 and HERV30 (Table 19).

Table 19. *Nucleotide identity between ERV1-1 genes and the related Endogenous Gammaretroviruses*

	ERV1-1 gag	***ERV1-1 pol***	***ERV1-1 env***
HERV-W	62	68	49
HERV9	62	64	4
HERV30	60	65	42

Pairwise similarity (%) between ERV1-1 and the related HERV-W, HERV9 and HERV30 *gag*, *pol* and *env* nucleotide sequence

We therefore further assessed ERV1-1 phylogenetic relationships with other endogenous and exogenous Gammaretroviruses [8, 292] at the amino acids level, by using Maximum

Likelihood (ML) analysis of Gag, Pol and Env putative protein (putein) sequences generated from the ERV1-1 proviral consensus sequence (Figure 20). To this aim, ERV1-1 proviral ORFs were predicted in Marmoset and Squirrel Monkey consensus sequences by ReTe [212], a software that attempts to reconstruct retroviral chains and the putative proteins encoded. Results showed that ERV1-1 Gag, Pol and Env puteins grouped together with HERV-W and other Gammaretrovirus-like families, such as HERV9 and HERV30, with a 100 bootstrap support (Figure 20).

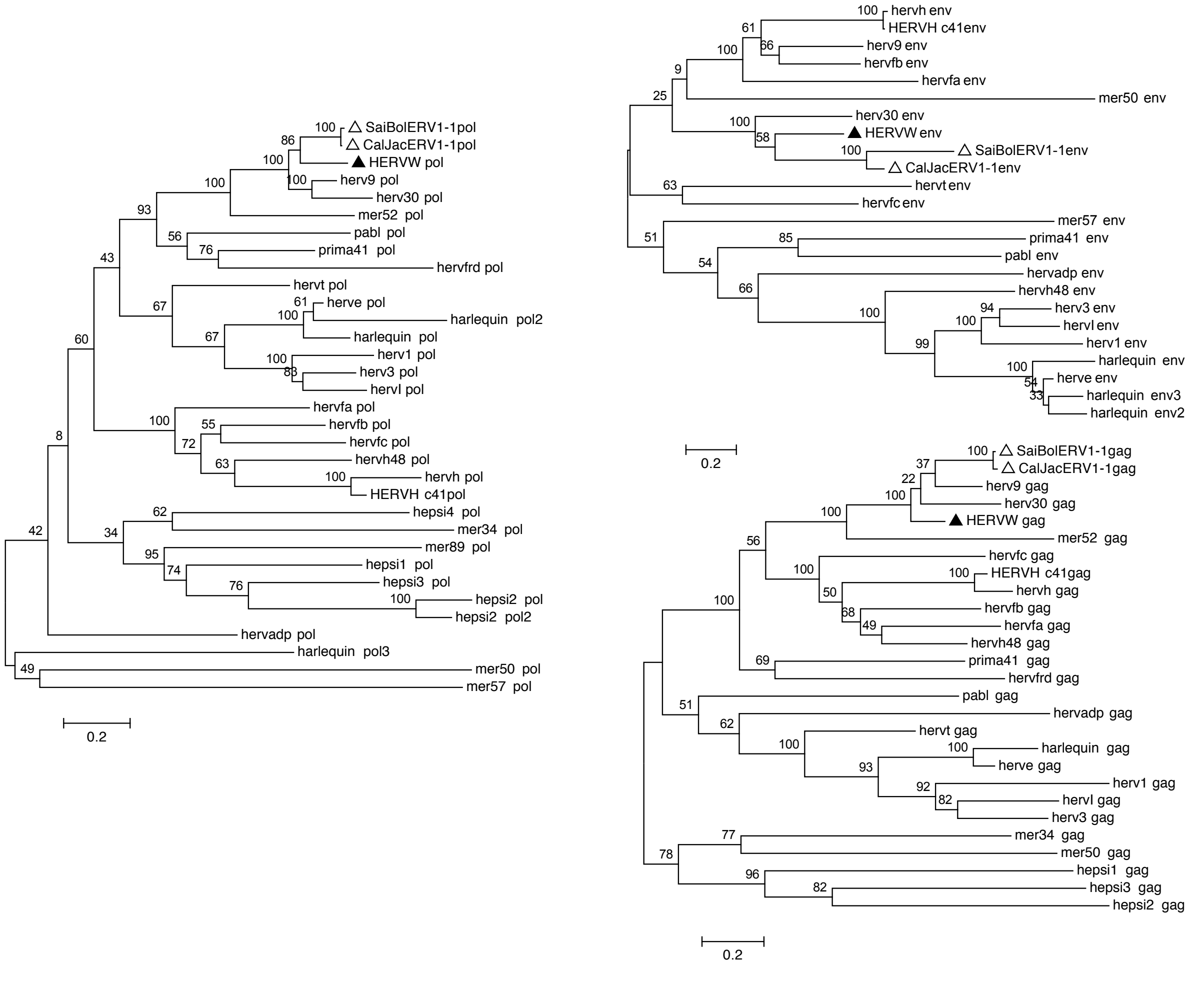


Figure 20. *Results from phylogenetic analysis of ERV1-1 Gag, Pol and Env and other Gammaretroviral protein and putein sequences.*
ERV1-1 puteins, labeled with an empty triangle, were obtained by identification and conceptual translation of Marmoset ERV1-1 proviral consensus sequence Open Reading Frames. That analysis was based on annotations for the ERV1-1_CJa-I reference sequence as provided by RepBase and output from RetroTector on-line [230]. The other Gammaretroviral putein sequences were from Vargiu et al. (2016). HERV-W puteins are marked with a filled triangle. The evolutionary relationships were inferred by using the Maximum Likelihood method based on the Poisson model. Phylogenies were tested by using the Bootstrap method with 100 replicates each. Length of branches indicates the number of substitutions per site.

In particular, for both Pol and Env ERV1-1 puteins, HERV-W was the most closely related sequence, further suggesting a strong evolutionarily relationship between the two groups, while such relation appeared to be less stringent considering the Gag putein. However, ERV1-1 Gag still was among the best RetroTector AutoFrame hits for HERV-W Gag recognition [8]. The same results were confirmed also in a NJ analysis with 1000 bootstrap replicates (data not shown). It is interesting to note that, even if HERV-W appears to be the closer relative to ERV1-1, also HERV9 and HERV30, already known as related to HERV-W, showed an evolutionary connection to ERV1-1, possibly suggesting a common origin of these Gammaretroviral HERVs.

4.8 Phylogeny and ERV1-1 sequence relationships with human solitary LTRs and HERV-W-like elements less similar to HERV17 and with *Catarrhini* ERV-W elements without human orthologs

In order to further characterize the ERV-W/HERV-W-like elements identified by BLAT searches of *Catarrhini* primate genome sequences using an LTR17-HERV17-LTR17 RepBase sequence as a query, the above mentioned three sets of sequences were compared with the consensus sequences of HERV-W, ERV1-1 and RepBase-derived sequences of other Human Endogenous Gammaretroviruses (the latter referred to as γHERVs).

4.8.1 ERV-W BLAT-identified sequences being solitary LTRs in human.

Alignment of human solitary LTRs, which constitute remnants of proviral insertions occurred in Rhesus (14) and Gibbon (5), with the sequences of other γHERVs confirmed that those solitary LTR belong to HERV-W group, as they clustered with the LTR17 consensus sequence with maximum bootstrap support (100), being clearly separated from all the other γHERVs (Figure 21). Furthermore, although HERV-W LTR17 and ERV1-1 LTR are very different in sequence, they nevertheless are closest relatives compared to other LTR sequences included in the analysis, even if showing a very low bootstrap value (14).

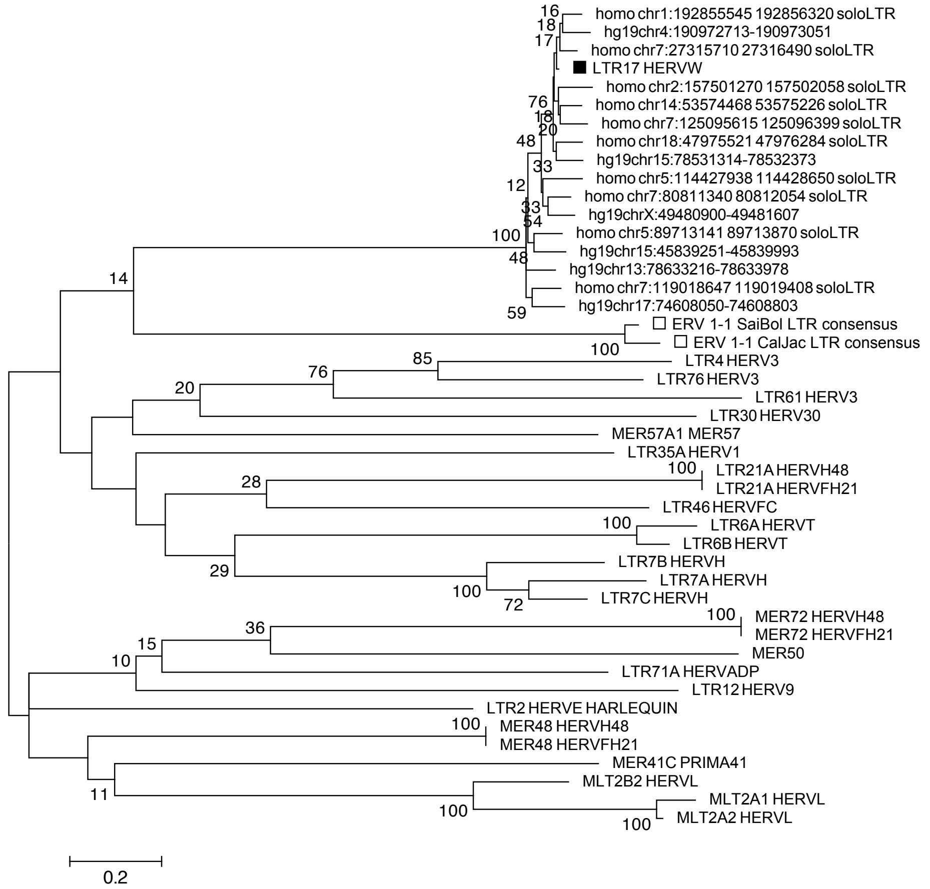

homo chr1:192855545 192856320 soloLTR
hg19chr4:190972713-190973051
homo chr7:27315710 27316490 soloLTR
■ LTR17 HERVW
homo chr2:157501270 157502058 soloLTR
homo chr14:53574468 53575226 soloLTR
homo chr7:125095615 125096399 soloLTR
homo chr18:47975521 47976284 soloLTR
hg19chr15:78531314-78532373
homo chr5:114427938 114428650 soloLTR
homo chr7:80811340 80812054 soloLTR
hg19chrX:49480900-49481607
homo chr5:89713141 89713870 soloLTR
hg19chr15:45839251-45839993
hg19chr13:78633216-78633978
homo chr7:119018647 119019408 soloLTR
hg19chr17:74608050-74608803
□ ERV 1-1 SaiBol LTR consensus
□ ERV 1-1 CalJac LTR consensus
LTR4 HERV3
LTR76 HERV3
LTR61 HERV3
LTR30 HERV30
MER57A1 MER57
LTR35A HERV1
LTR21A HERVH48
LTR21A HERVFH21
LTR46 HERVFC
LTR6A HERVT
LTR6B HERVT
LTR7B HERVH
LTR7A HERVH
LTR7C HERVH
MER72 HERVH48
MER72 HERVFH21
MER50
LTR71A HERVADP
LTR12 HERV9
LTR2 HERVE HARLEQUIN
MER48 HERVH48
MER48 HERVFH21
MER41C PRIMA41
MLT2B2 HERVL
MLT2A1 HERVL
MLT2A2 HERVL
16
18
17
76
18
20
48
33
12
33
54
100
48
59
14
100
85
76
20
100
28
100
29
100
72
100
36
15
10
100
11
100
100
0.2

Figure 21. *Result from phylogenetic analysis of nucleotide sequences of human solitary LTRs representing orthologous loci initially formed in common ancestors of human with Rhesus or Gibbon.*
Endogenous Gammaretroviral LTR sequences were retrieved from RepBase.. The HERV-W group LTR17 reference sequence is marked with a filled square. The ERV1-1 LTR consensus, generated from the Marmoset (CalJac) and Squirrel Monkey (SaiBol) proviral sequence datasets, and are marked with empty squares. Evolutionary relationships were inferred by using the Maximum Likelihood method and the Kimura-2-parameter model. The resulting phylogeny was tested by using the Bootstrap method with 100 replicates. Length of branches indicates the number of substitutions per site.

4.8.2 ERV-W BLAT-identified sequences corresponding to HERV-W-like elements with lesser identity to HERV17

We next assessed phylogenetic relationships of human HERV-W-like elements showing lesser sequence identity to HERV17 with HERV-W and ERV1-1 consensus sequences and other γHERV sequences (Figure 22).

Three clusters of HERV-W-like sequences with good or reasonable bootstrap support could be defined: cluster I, 96%; cluster II, 100%; cluster III, 70%. The three clusters were separated from the other γHERV sequences with a 96% bootstrap support and included 24 out of 29 HERV-W-like sequences as well as HERV-W, HERV9, HERV30 and ERV1-1. HERV-W-like elements in cluster I were most related to HERV-W, while HERV-W-like elements in cluster II were more closely related to HERV9 and HERV30. In this particular analysis, ERV1-1 was most closely related to HERV30 and HERV9 (Figure 22). RepeatMasker analysis (Table 17) provided further support in that cluster I members were annotated exclusively as HERV17, while cluster II members include elements structurally related to HERV17 and, in one case, HERV30 when regarding the proviral internal portions, yet they harbor LTR12F as LTR type (the HERV9 LTR in RepBase).

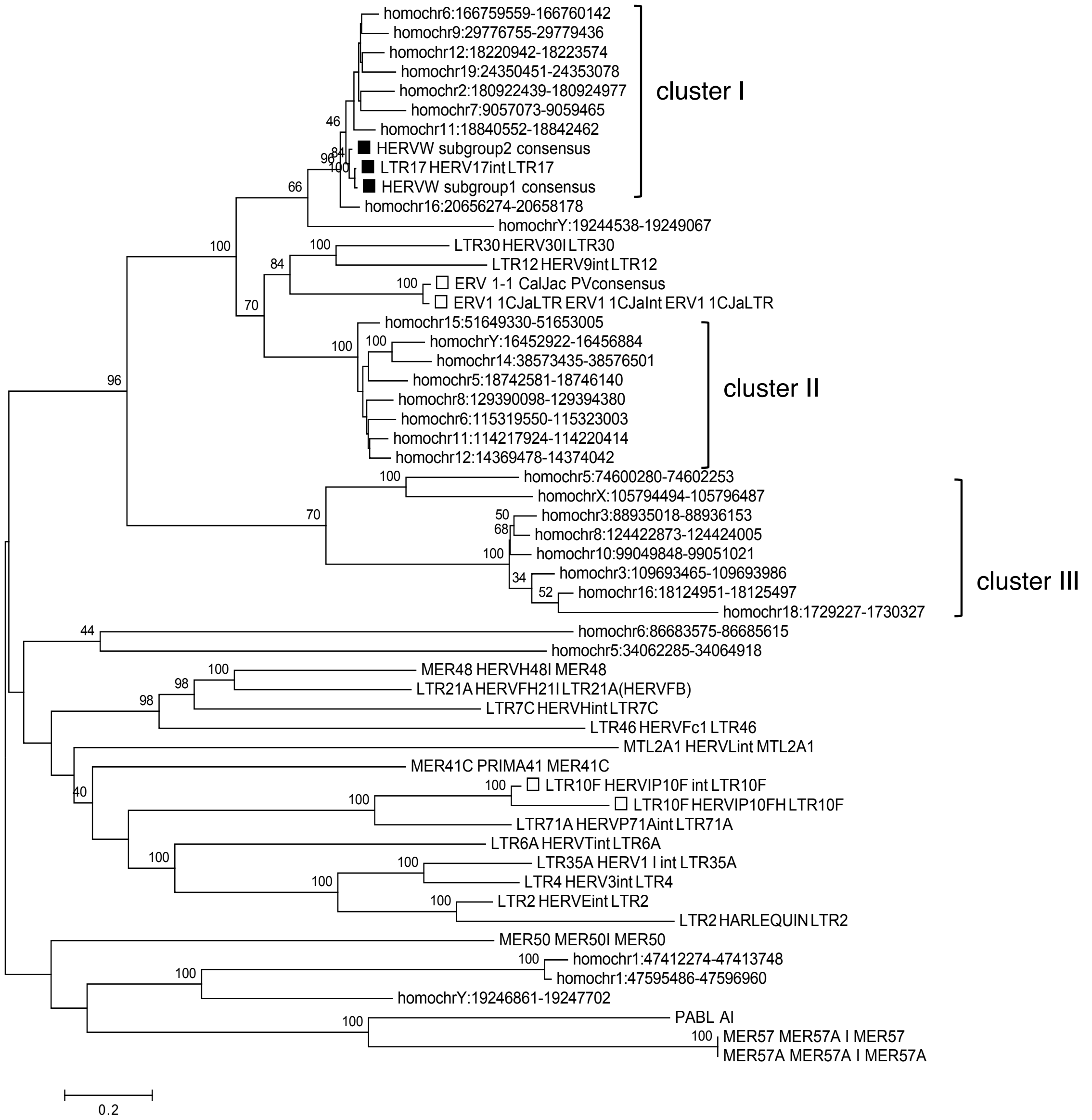

homochr6:166759559-166760142
homochr9:29776755-29779436
homochr12:18220942-18223574
homochr19:24350451-24353078
homochr2:180922439-180924977
homochr7:9057073-9059465
46
homochr11:18840552-18842462
84
■ HERVW subgroup2 consensus
96
100
■ LTR17 HERV17int LTR17
■ HERVW subgroup1 consensus
66
homochr16:20656274-20658178
cluster I
homochrY:19244538-19249067
100
100
LTR30 HERV30I LTR30
84
LTR12 HERV9int LTR12
100
□ ERV 1-1 CalJac PVconsensus
70
□ ERV1 1CJaLTR ERV1 1CJaInt ERV1 1CJaLTR
homochr15:51649330-51653005
100
100
homochrY:16452922-16456884
homochr14:38573435-38576501
96
homochr5:18742581-18746140
homochr8:129390098-129394380
homochr6:115319550-115323003
homochr11:114217924-114220414
homochr12:14369478-14374042
cluster II
100
homochr5:74600280-74602253
homochrX:105794494-105796487
70
50
homochr3:88935018-88936153
68
homochr8:124422873-124424005
100
homochr10:99049848-99051021
34
homochr3:109693465-109693986
52
homochr16:18124951-18125497
homochr18:1729227-1730327
cluster III
44
homochr6:86683575-86685615
homochr5:34062285-34064918
100
MER48 HERVH48I MER48
98
LTR21A HERVFH21I LTR21A(HERVFB)
98
LTR7C HERVHint LTR7C
LTR46 HERVFc1 LTR46
MTL2A1 HERVLint MTL2A1
MER41C PRIMA41 MER41C
40
100
□ LTR10F HERVIP10F int LTR10F
100
□ LTR10F HERVIP10FH LTR10F
LTR71A HERVP71Aint LTR71A
LTR6A HERVTint LTR6A
100
100
LTR35A HERV1 I int LTR35A
100
LTR4 HERV3int LTR4
100
LTR2 HERVEint LTR2
LTR2 HARLEQUIN LTR2
MER50 MER50I MER50
100
homochr1:47412274-47413748
100
homochr1:47595486-47596960
homochrY:19246861-19247702
100
PABL AI
100
MER57 MER57A I MER57
MER57A MER57A I MER57A
0.2

Figure 22*. Result from phylogenetic analysis of HERV-W-like nucleotide sequences orthologous to ERV-W loci identified in non-human primates by HERV17 BLAT searches.*

Endogenous Gammaretroviral sequences were retrieved from RepBase. The HERV-W group LTR17 reference sequence from RepBase and the proviral HERV-W LTR consensus sequence generated previously (Chapter 3, Grandi et al. 2016), and are marked with a filled square. The ERV1-1 reference sequence from RepBase and the proviral consensus generated from the proviral sequence dataset generated in this study are marked with an empty square. Evolutionary relationships were inferred by using the Maximum Likelihood method and the Kimura-2-parameter model. The resulting phylogeny was tested by using the Bootstrap method with 100 replicates. Length of branches indicates the number of substitutions per site.

The remaining 12 cluster III sequences were only remotely related to the other HERV-W-like elements (bootstrap support = 52), with that branch including sequences identified either as LTR17-HERV17 or related to other γHERVs (HERV9, HERV30, HERVH, HERVIP10FH) based on RepeatMasker analysis (Table 17). However, such sequence relationships were not confirmed by phylogenetic analysis, in which cluster III members were clearly separated from γHERVs consensus sequences (Figure 22). Overall, this analysis demonstrated closer relationships, of yet different degrees, of human HERV-W-like elements with HERV-W, HERV9, HERV30 and ERV1-1.

4.8.3 ERV-W BLAT-identified sequences lacking an orthologous in humans

To verify the relationship of *Catarrhini* ERV-W sequences lacking an orthologs in human to the ERV-W group, respective sequences were compared with HERV-W and other γHERV sequences. Chimp, Gorilla, Orangutan and Gibbon full-length sequences were analyzed separately (Figure 23) from the Rhesus-derived ERV-W sequences because of the relatively high number of Rhesus sequences. The latter phylogeny, for the same reason, was inferred based on the *pol* gene instead of the full-length sequence (Figure 24).

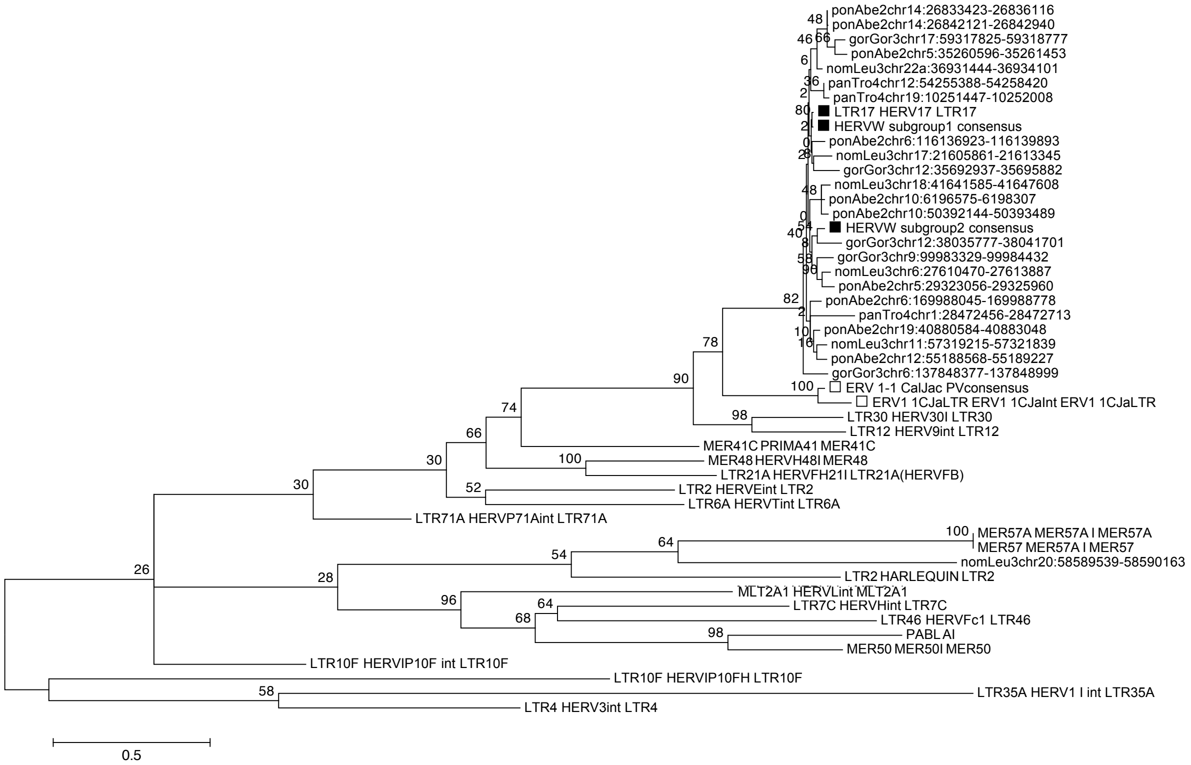

48
ponAbe2chr14:26833423-26836116
ponAbe2chr14:26842121-26842940
46
66
gorGor3chr17:59317825-59318777
ponAbe2chr5:35260596-35261453
6
nomLeu3chr22a:36931444-36934101
36
panTro4chr12:54255388-54258420
2
panTro4chr19:10251447-10252008
80
■ LTR17 HERV17 LTR17
2
■ HERVW subgroup1 consensus
0
ponAbe2chr6:116136923-116139893
28
nomLeu3chr17:21605861-21613345
gorGor3chr12:35692937-35695882
48
nomLeu3chr18:41641585-41647608
ponAbe2chr10:6196575-6198307
0
ponAbe2chr10:50392144-50393489
54
■ HERVW subgroup2 consensus
40
8
gorGor3chr12:38035777-38041701
58
gorGor3chr9:99983329-99984432
90
nomLeu3chr6:27610470-27613887
ponAbe2chr5:29323056-29325960
82
ponAbe2chr6:169988045-169988778
2
panTro4chr1:28472456-28472713
10
ponAbe2chr19:40880584-40883048
16
nomLeu3chr11:57319215-57321839
ponAbe2chr12:55188568-55189227
78
gorGor3chr6:137848377-137848999
90
100
□ ERV 1-1 CalJac PVconsensus
□ ERV1 1CJaLTR ERV1 1CJaInt ERV1 1CJaLTR
74
98
LTR30 HERV30I LTR30
LTR12 HERV9int LTR12
66
MER41C PRIMA41 MER41C
30
100
MER48 HERVH48I MER48
LTR21A HERVFH21I LTR21A(HERVFB)
30
52
LTR2 HERVEint LTR2
LTR6A HERVTint LTR6A
LTR71A HERVP71Aint LTR71A
100
MER57A MER57A I MER57A
MER57 MER57A I MER57
64
nomLeu3chr20:58589539-58590163
54
26
28
LTR2 HARLEQUIN LTR2
MLT2A1 HERVLint MLT2A1
96
64
LTR7C HERVHint LTR7C
68
LTR46 HERVFc1 LTR46
98
PABL AI
MER50 MER50I MER50
LTR10F HERVIP10F int LTR10F
LTR10F HERVIP10FH LTR10F
58
LTR35A HERV1 I int LTR35A
LTR4 HERV3int LTR4
0.5

Figure 23. *Result from phylogenetic analysis of Chimpanzee, Gorilla, Orangutan and Gibbon ERV-W nucleotide sequences lacking an orthologous in the human genome sequence.*
Endogenous Gammaretroviral sequences were retrieved from RepBase. The HERV17/HERV-W group reference sequence was retrieved from RepBase and the proviral consensus sequences were from the proviral dataset

generated previously (Chapter 3, Grandi et al. 2016), and are marked with a filled square. The ERV1-1 reference sequence from RepBase and the proviral consensus generated from the proviral sequence dataset generated in this study are marked with an empty square. Evolutionary relationships were inferred by using the Maximum Likelihood method and theKimura-2-parameter model. The resulting phylogeny was tested by using the Bootstrap method with 100 replicates. Length of branches indicates the number of substitutions per site.

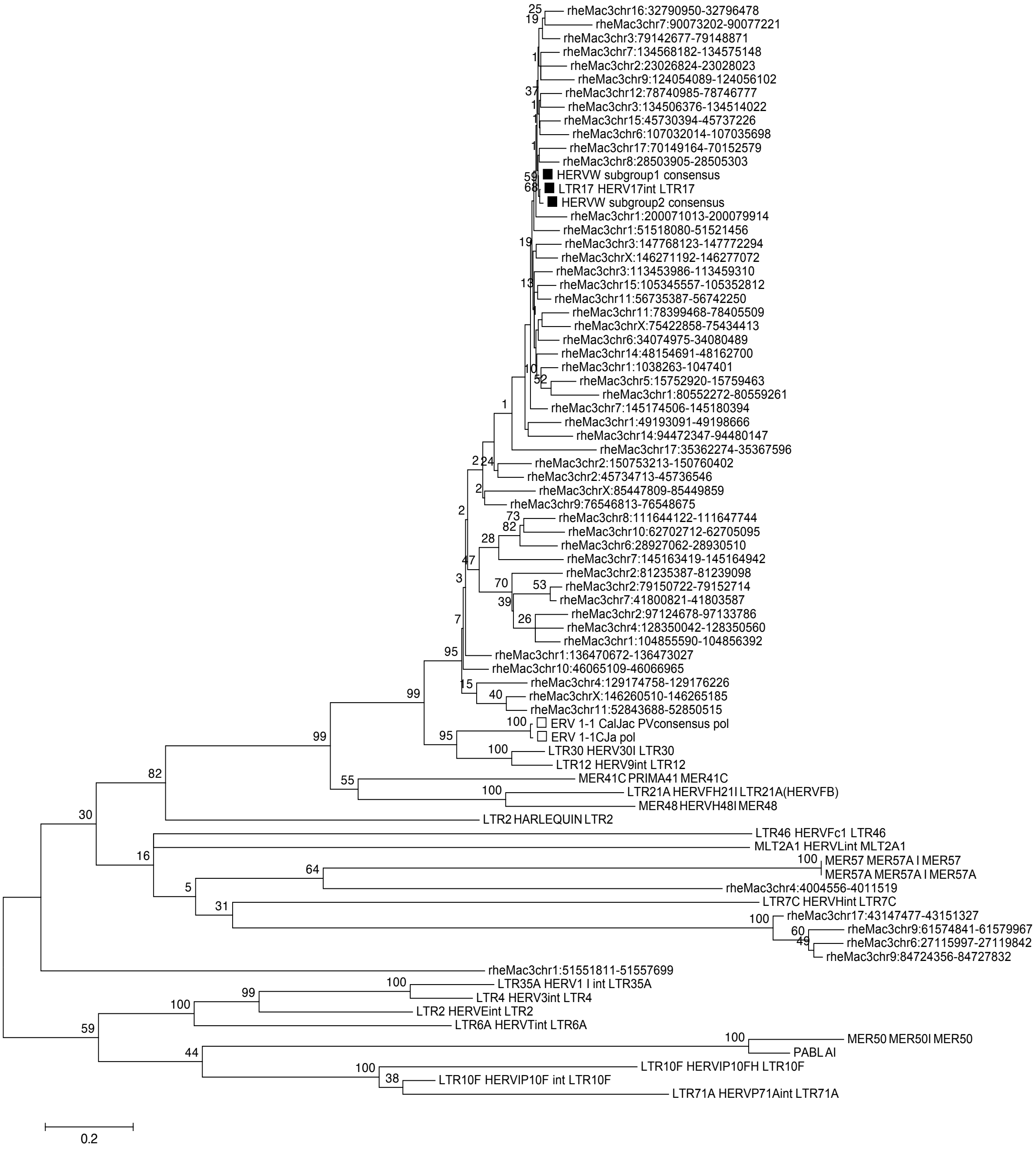

25
19
rheMac3chr16:32790950-32796478
rheMac3chr7:90073202-90077221
rheMac3chr3:79142677-79148871
1
rheMac3chr7:134568182-134575148
rheMac3chr2:23026824-23028023
rheMac3chr9:124054089-124056102
37
rheMac3chr12:78740985-78746777
1
rheMac3chr3:134506376-134514022
1
rheMac3chr15:45730394-45737226
rheMac3chr6:107032014-107035698
1
rheMac3chr17:70149164-70152579
rheMac3chr8:28503905-28505303
59
■ HERVW subgroup1 consensus
68
■ LTR17 HERV17int LTR17
■ HERVW subgroup2 consensus
rheMac3chr1:200071013-200079914
rheMac3chr1:51518080-51521456
19
rheMac3chr3:147768123-147772294
rheMac3chrX:146271192-146277072
rheMac3chr3:113453986-113459310
13
rheMac3chr15:105345557-105352812
rheMac3chr11:56735387-56742250
1
rheMac3chr11:78399468-78405509
rheMac3chrX:75422858-75434413
rheMac3chr6:34074975-34080489
rheMac3chr14:48154691-48162700
10
rheMac3chr1:1038263-1047401
52
rheMac3chr5:15752920-15759463
rheMac3chr1:80552272-80559261
1
rheMac3chr7:145174506-145180394
rheMac3chr1:49193091-49198666
rheMac3chr14:94472347-94480147
rheMac3chr17:35362274-35367596
2
24
rheMac3chr2:150753213-150760402
rheMac3chr2:45734713-45736546
2
rheMac3chrX:85447809-85449859
rheMac3chr9:76546813-76548675
2
73
rheMac3chr8:111644122-111647744
82
rheMac3chr10:62702712-62705095
28
rheMac3chr6:28927062-28930510
47
rheMac3chr7:145163419-145164942
3
rheMac3chr2:81235387-81239098
70
53
rheMac3chr2:79150722-79152714
39
rheMac3chr7:41800821-41803587
7
26
rheMac3chr2:97124678-97133786
rheMac3chr4:128350042-128350560
rheMac3chr1:104855590-104856392
95
rheMac3chr1:136470672-136473027
rheMac3chr10:46065109-46066965
99
15
rheMac3chr4:129174758-129176226
40
rheMac3chrX:146260510-146265185
rheMac3chr11:52843688-52850515
99
100
□ ERV 1-1 CalJac PVconsensus pol
95
□ ERV 1-1CJa pol
100
LTR30 HERV30I LTR30
LTR12 HERV9int LTR12
82
55
MER41C PRIMA41 MER41C
100
LTR21A HERVFH21I LTR21A(HERVFB)
MER48 HERVH48I MER48
30
LTR2 HARLEQUIN LTR2
LTR46 HERVFc1 LTR46
16
MLT2A1 HERVLint MLT2A1
64
100
MER57 MER57A I MER57
MER57A MER57A I MER57A
5
rheMac3chr4:4004556-4011519
31
LTR7C HERVHint LTR7C
100
rheMac3chr17:43147477-43151327
60
rheMac3chr9:61574841-61579967
49
rheMac3chr6:27115997-27119842
rheMac3chr9:84724356-84727832
rheMac3chr1:51551811-51557699
100
LTR35A HERV1 I int LTR35A
99
LTR4 HERV3int LTR4
100
LTR2 HERVEint LTR2
59
LTR6A HERVTint LTR6A
100
MER50 MER50I MER50
44
PABL AI
100
LTR10F HERVIP10FH LTR10F
38
LTR10F HERVIP10F int LTR10F
LTR71A HERVP71Aint LTR71A
0.2

Figure 24. *Result from phylogenetic analysis of Rhesus ERV-W locus nucleotide sequences lacking an orthologous in the human reference genome sequence.*

Endogenous Gammaretroviral consensus sequences were retrieved from RepBase. The HERV17/HERV-W group reference sequence was retrieved from RepBase and the other HERV-W proviral consensus sequences were from the dataset generated previously [121], and are marked with a filled square. The ERV1-1 reference sequence from RepBase and the proviral consensus generated from the proviral sequence dataset generated in this study are marked with an empty square. Evolutionary relationships among the pol gene nt sequences were inferred by using the Maximum Likelihood method and the Kimura-2-parameter model. The resulting phylogeny was tested by using the Bootstrap method with 100 replicates. Length of branches indicates the number of substitutions per site.

All ERV-W sequences without orthologs in the human genome sequence identified in *Catarrhini* species Chimpanzee, Gorilla, Orangutan and Gibbon grouped with the HERV-W consensus sequence (bootstrap 82%) and were furthermore closely related to ERV1-1 (bootstrap 78%) followed by HERV9 and HERV30 (Figure 23). A single sequence retrieved from Gibbon (nomLeu3; chr20:58,589,539-58,590,163) displayed a, yet, weakly supported (64%) relationship with MER57.

The separately analyzed Rhesus-derived ERV-W sequences lacking an orthologous in human formed a well-supported (90%) cluster with HERV-W, further corroborating the ERV-W nature of those sequences (Figure 24). That phylogenetic clade was likewise related to HERV9 and HERV30 in a very similar fashion and with high bootstrap supports (99%). Six Rhesus ERV-W sequences located outside of that cluster (Figure 24). In particular, 1 sequence was more related to MER57 (bootstrap support 64%), 4 sequences clustered together with 100% bootstrap support and were weakly related to the HERVH consensus sequence, yet with poor bootstrap support of 31%, while 1 sequence formed a separate branch.

To further examine the actual nature of those 6 Rhesus-derived sequences, we compared their full-length nucleotide sequences to a subset of γHERVs reference sequences by EMBOSS polydot analysis (Figure 25).

The sequence in Rhesus chr4:4,004,556-4,011,519 shares longer stretches of nucleotide identity exclusively with the HERV-W consensus sequence, while the 4 Rhesus sequences weakly related to HERV-H, based on *pol* phylogeny (Figure 8), displayed longer stretches of sequence similarity with both the HERV-W and HERV-H consensus sequences, thus possibly representing non-canonical recombinants. In contrast, the sequence in Rhesus chr1:51,551,811-51,557,699, forming a separate branch in a *pol* gene tree, did not show appreciable similarity to any of the γHERV consensus sequences.

Taken together, phylogenetic analysis confirmed the ERV-W nature of almost all the retrieved ERV-W-like elements without human orthologs in non-human *Catarrhini* species as well as the independent spread of "true" (H)ERV-W elements in Rhesus later in primate evolution.

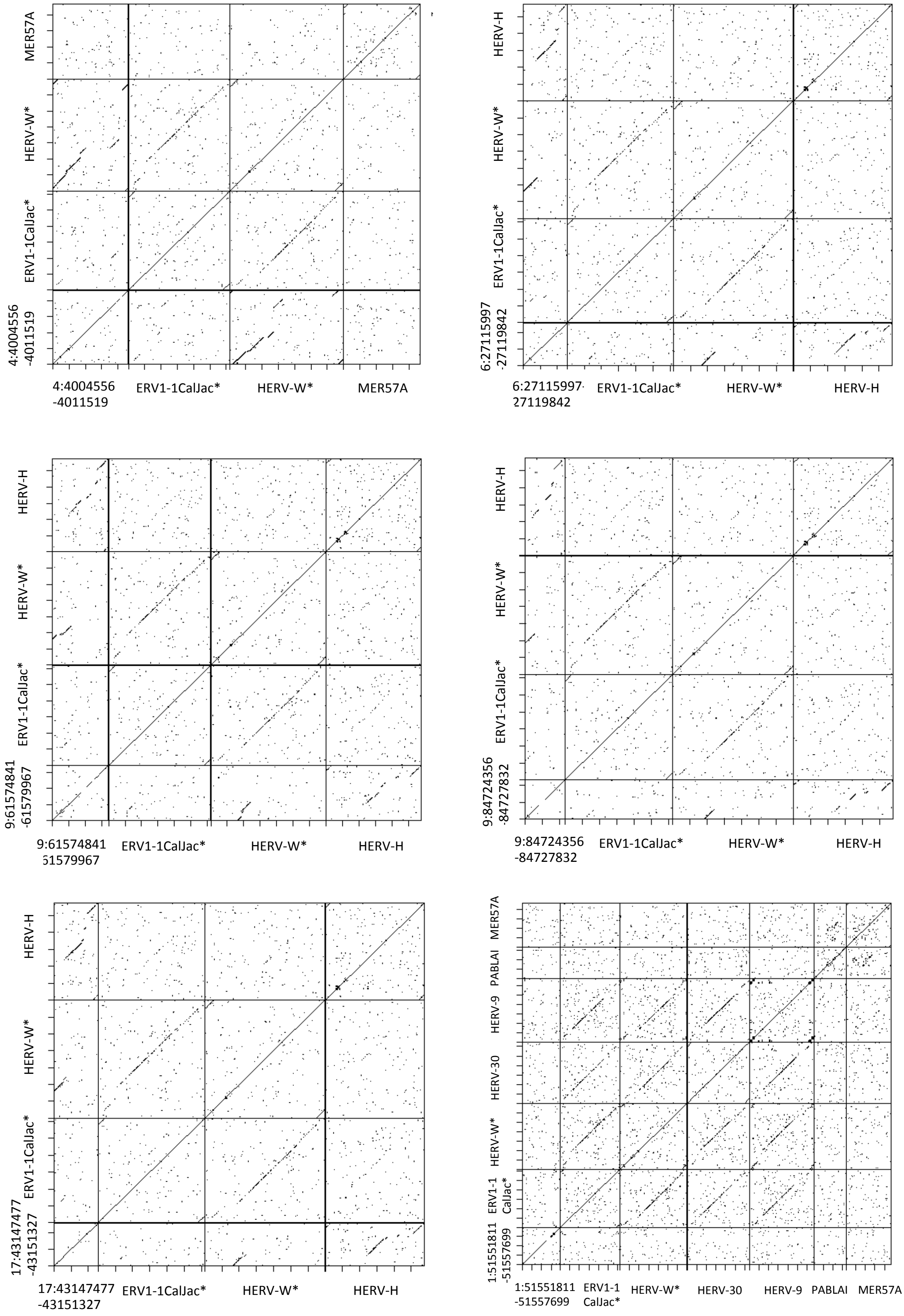


***Figure 25**. Polydot pairwise sequence comparisons of the 6 Rhesus ERV-W nucleotide sequences lacking an orthologous in the human reference genome sequence and showing unclear sequence relationships with endogenous Gammaretroviral HERV sequences.* Analyzed consensus sequences marked with "*" were generated in this study. Other sequences were retrieved from RepBase.

4.9 Discussion

Following up on our recent detailed characterization of the HERV-W group in the human genome (Chapter 3 and [21]), the present work was aimed at analyzing in more detail ERV-W elements integrated in genome sequences of non-human primates, to provide a better depiction of the diffusion of ERV-W during the evolution of primates, thus contributing insights into only partially resolved aspects. For instance, the initial entry of ERV-W elements into the primate lineage is currently not fully resolved. A number of studies suggested that the ERV-W group's initial germ line colonization had occurred in *Catarrhini* primates after the lineage's evolutionary separation from *Platyrrhini* primates, i.e. less than ~ 40 MYa, being based either on HERV-W *pol* PCR [255] or Southern Blot [256] analysis in different Old World Monkeys samples, or on the nucleotide divergence calculation between HERV-W subfamilies [34]. Such results were supported also by the absence of the same ERV-W sequences in *Platyrrhini* and *Prosimians* [34, 255, 256]. One of these works reported, in addition, the presence of the sole ERV-W LTRs not only limited to *Catarrhini* primates, but also in the three *Platyrrhini* species analyzed by PCR amplification, suggesting an acquisition occurred approximately 55 million years ago [255]. Overall, these studies gave thus just a general rough overview of HERV-W group acquisition by primates genomes, without information about the single members diffusion and the precise time period of activity of the group. Beside the single finding of ERV-W LTRs presence in *Platyrrhini* [255], the available information together suggest that the first (H)ERV-W proviral acquisitions have occurred around 25 MYa, and the group as a whole have formed during a rather short period of proliferation (~ 5 MY) [34, 36, 256]. Such relatively low proliferation rate has been explained by the presence of two types of HERV-W sequences: proviruses and L1-mediated processed pseudogenes: the latter miss the 5′LTR U3 and 3'LTR U5 regions, thus being proliferation-incompetent [34].

Our analysis provided further support that the ERV-W group is present exclusively in *Catarrhini* primates, but the search for ERV-W orthologous loci in the genomes of Hominoids and OWMs revealed that the group was instead proliferating for a prolonged time period, with novel locus formations having occurred approximately between 43 and 20 MYa, in line with recent age estimation of single HERV-W sequences (Chapter 3 and ref. [21]). Interestingly, a 2:1 ratio of L1-mediated processed pseudogenes formations relative to "true" provirus formations was observed in Rhesus and Gibbon, suggesting that a quite massive formation of ERV-W processed pseudogenes occurred in parallel to formation of "true" proviruses. Similarly, also in Orangutan and Gorilla, ERV-W processed pseudogenes were the main source of additional ERV-W loci acquisition.

The spread of the ERV-W group within *Catarrhini* parvorder was further investigated by the identification of ERV-W elements in non-human primates genome sequences through BLAT searches at the UCSC Genome Browser, using RepBase HERV17 reference sequence as a query. That strategy identified 4 ERV-W loci in Gibbon and 15 in Rhesus that were likely formed between 43 and 20 MYa and were present in the human genome only as solitary LTR each. In support of a longer time period of ERV-W locus formations, some ERV-W loci in non-human primates appeared to have formed species-specifically and thus lack orthologs in the respective other primates and especially in the human genome. In particular, we identified 88 ERV-W loci in non-human primates with corresponding empty sites in the human genome, 81 of which could be interpreted as species-specific insertions in respective primates: 66 in Rhesus, 2 in Gibbon, 6 in Orangutan, 4 in Gorilla, and 3 in Chimpanzee, the latter further indicating lineage-specific formation of ERV-W loci less than 10 MYa. Importantly, species-specific formation of ERV-W loci occurred by both by formation of "true" proviruses insertion and L1-mediated processed pseudogenes.

BLAT searches furthermore identified 29 ERV-W-like elements with somewhat lesser similarities to HERV-W, mostly present in the Rhesus genome, yet some of them also found in Gibbon (3), Orangutan (2) and Gorilla (2). Of further note, our analysis could not confirm presence of ERV-W sequences in NWMs regarding neither proviral gene regions nor LTRs.

It should be stressed here that our analysis of (orthologous) ERV-W loci present (or absent) in the various available primate genome sequences relies on primate comparative genomics data as provided by the UCSC Genome Browser [214, 216] that were generated by (b)lastz alignments [291] of primate genome sequence data. Furthermore, our own analysis of orthologous loci required a minimum of 500 nt of upstream and downstream flanking sequence to possible ensure analysis of truly homologous genome regions in the different primate genome sequences. While some of the observed differences in orthologous ERV-W loci may be due to errors in genome sequence assemblies or (b)lastz alignments, it appears that only a minority of loci are associated with, or in close proximity to, for instance, gaps in assembled genome sequences.

Taken together, our comparative analysis of primate genome sequences thus provides a more detailed evolutionary history of (H)ERV-W and the spread of ERV-W during *Catarrhini* primate evolution, corroborating an extended period of ERV-W locus formations having peaked between approximately 42 and 30 MYa and providing sporadic, species or lineage-specific ERV-W locus formations until less than 10 MYa.

Of further note, our sequence searches corroborated an ERV group, named ERV1-1_CJa in Repbase, as closely related to ERV-W. Because of the lack of an established ERV nomenclature, we designated those sequences as ERV1-1. A total of 130 ERV1-1 loci were identified in the genomes of Marmoset (59) and Squirrel Monkey (71). Sequence searches of unassembled genome sequence data furthermore indicated that ERV1-1 is present in the three *Platyrrhini* families. However, there was no evidence of ERV1-1 sequences in genome sequence of both *Tarsiiformes* and *Prosimians*, indicating that ERV1-1 sequences were formed in the respective primate lineage less than 60 Mya based on estimated times of separations of respective primate lineages [221, 222]. Also noteworthy, none of the ERV1-1 loci showed signatures of processed pseudogenes generated by L1, as it is the case for many (H)ERV-W loci [33, 34]. The established close sequence relationships at both the nucleotide and amino acid level suggest that ERV1-1 and (H)ERV-W derive from a common ancestor also involving related groups HERV9 and HERV30. Such closer sequence relationships do not apply to the ERV1-1 proviral LTRs that are very different in sequence from (H)ERV-W LTRs. This is in line with previous observations for some ERV groups in which different evolutionary paths of sequence evolution were taken by the proviral body and associated LTR sequences, resulting in different LTR subgroups associated with otherwise monophyletic proviral bodies (for instance, see [295, 296]) and possibly leading to the formation of retroviral chimeras [8].

In conclusion, the present study provides an exhaustive overview of the germ line colonization of ERV-W during the evolution of primates. Our study revealed a rather unexpectedly long period of ERV-W activity, provided evidences of separate activation of ERV-W, especially in the Rhesus evolutionary lineage, and pointed out that L1-mediated formation of ERV-W processed pseudogenes was not a secondary phenomenon with negative impact on the group's proliferation rate, but instead a parallel and major mechanism of ERV-W locus formations in all primates genomes.

Chapter 5. Characterization of the ERV1-1 sequences in Marmoset and Squirrel Monkey (Platyrrhini parvorder): further insights into the group close phylogenetic relation to ERV-W

5.1 Introduction

As reported in Chapter 4, the Genome Browser BLAT search for ERV-W elements in Marmoset and Squirrel Monkey *Platyrrhini* genome assemblies led to retrieve sequences showing high nucleotide identity to HERV17, previously annotated in the respective Genome Browser reference sequence as ERV1-1_CJa-I for the internal portion and ERV1-1_CJa-LTR for the 5′ and 3′ LTRs. In particular, ERV-W and ERV1-1 groups showed a remarkable similarity regarding the retroviral internal portion, and the close identity between the two ERV groups was confirmed by a strong relationship at the phylogenetic level, shared also with some other HERV-W related Gammaretroviral ERVs (HERV30, HERV9) (Chapter 4, Figures 19 and 20, Table 19). These results allowed to characterize, for the first time, the close phylogenetic relationship between the ERV1-1 group in *Platyrrhini* primates and the ERV-W elements in *Catarrhini* species genomes, possibly suggesting the presence of a common ancestor between these groups and shared also by related ERV Gammaretroviruses.

At our best knowledge, ERV1-1 elements integrated within vertebrates genomes are poorly characterized, and no information is available about the group presence in primates, with the exception of the RepBase entries for ERV1-1 sequences retrieved in Marmoset (*Simiiformes, Platyrrhini* parvorder) (ERV1-1_CJa) and Tarsier (*Tarsiiformes*) (ERV1-1_TSy).

Hence, in order to have more insights on the evolutionarily connection between ERV-W and ERV1-1 groups, a total of 130 ERV1-1 most complete proviral sequences were retrieved from Marmoset and Squirrel Monkey *Platyrrhini* genomes and analyzed for their main structural elements, showing various features in common with ERV-W. Moreover, similarly to ERV-W elements, the LTRs phylogenetic analysis allowed to classify ERV1-1 sequences into 4 main subgroups, supported by bootstrap values ≥ 90 each. The estimation of the single ERV1-1 sequences time of integration pointed out different period of diffusion for the various ERV1-1 subgroups, with a general entry in primates lineage occurred before the ERV-W diffusion.

5.2 ERV1-1 groups presence in vertebrates genomes

In Chapter 4, we introduced a group of ERV-W related elements retrieved by HERV17 BLAT search in the genome sequences of Marmoset and Squirrel Monkey (*Platyrrhini*), annotated in Genome Browser as ERV1-1_CJa-I for the internal portion and ERV1-1_CJa-LTR for LTRs, and subsequently indicated as ERV1-1. In RepBase Update database of repetitive elements [295], the ERV1-1 entry could be associated to ERV sequences present in at least 29 genomic assemblies of vertebrates species, including two primates (Marmoset and Tarsier). Hence, we firstly asked which relationship would be present among the ERV1-1 elements identified in the different vertebrates species, including in a phylogenetic tree each ERV1-1 RepBase consensus sequences for the retroviral internal portion and the HERV-W proviral consensus sequence obtained from the human dataset (Chapter 3 and ref. [21]). As shown in Figure 26, Marmoset ERV1-1 RepBase consensus sequence clustered in a phylogenetic clade supported by a 72 bootstrap (indicated with a red bracket), including most of the ERV1-1 groups integrated within mammals genomes, together with some reptiles and a fish ERV1-1 groups.

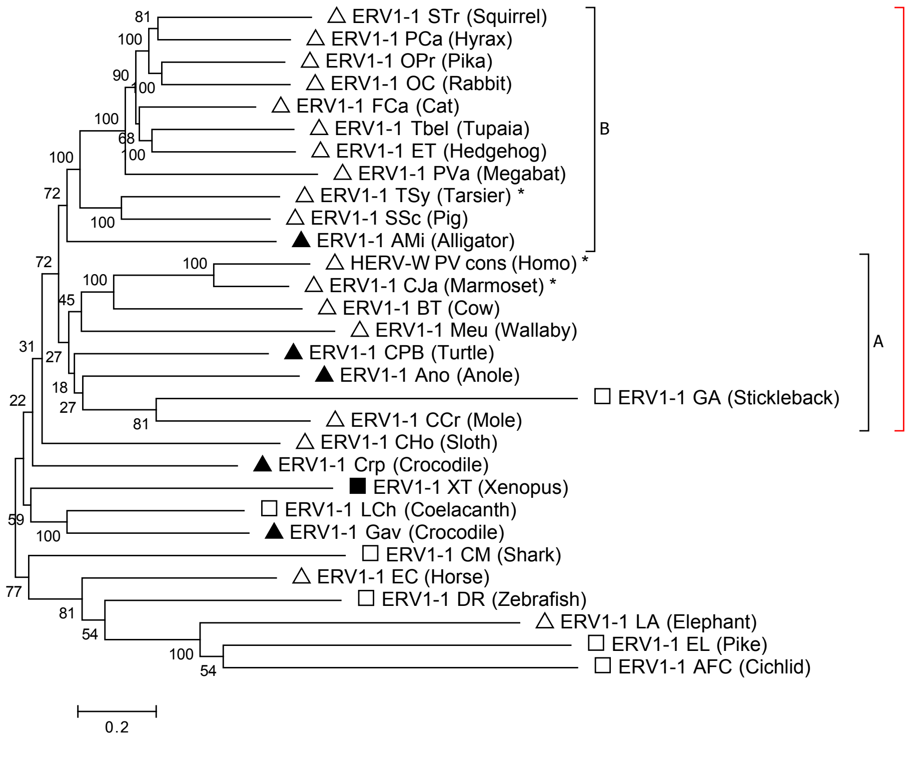


Figure 26. Phylogenetic relationships between the different RepBase ERV1-1 vertebrates groups (internal portion).

The consensus sequences representing the internal portion of the ERV1-1 sequences retrieved in various vertebrate genomes were collected from RepBase Update. Among the vertebrates species taken into account, empty triangles indicate mammals, filled triangles indicate reptiles, empty squares indicate fishes and filled squares indicate amphibians. The RepBase entries corresponding to ERV1-1 elements retrieved in primates genomes are further marked with an asterisk. Phylogenetic relationships were inferred on the full-length retroviral internal portion nucleotide sequence by using the Neighbor Joining method and the Kimura-2-parameter model and applying pairwise deletion option. The resulting phylogeny was tested by using the Bootstrap method with 1000 replicates. Length of branches indicates the number of substitutions per site.

In particular, Marmoset ERV1-1 consensus sequence shared the closest phylogenetic relation with the HERV-W proviral consensus sequence, supported by the highest bootstrap support (100), being also highly related to ERV1-1 elements in Cow *(Bos Taurus)*. HERV-W and ERV1-1 Marmoset and Bovine ERV1-1 consensus further clustered with other ERV1-1 elements found in mammals, reptiles and a fish species, forming a branch (indicated with the letter "A" in Figure 26) weakly supported by bootstrap values. The other RepBase ERV1-1 entry retrieved from a primate species, i.e. Tarsier, was also included in the same phylogenetic clade, being however more closely related to other ERV1-1 groups retrieved in mammals (and sequence from Alligator reptiles in one case) and forming thus a separate branch (indicated with the letter "B" in the figure) with bootstrap supports of 72 (mammals and Alligator ERV1-1 elements) and 100 (only mammals ERV1-1 sequences). Differently from the ERV1-1 groups forming the above-described two branches, a number of ERV1-1 vertebrates consensus sequences were not included in that 72 bootstrap supported phylogenetic clade, resulting instead lowly related to the latter. These low related ERV1-1 sequences mostly enclosed elements retrieved from fishes, showing also phylogenetic relations with some reptiles and mammals species ERV1-1 and with the sole ERV1-1 consensus sequence retrieved from amphibians.

The same analysis performed using a ML phylogenetic tree gave very similar results (data not shown).

Similarly to the ERV1-1 retroviral internal portion, also the ERV1-1 LTR consensus sequences listed in RepBase for the same vertebrates species (except for Stickleback) were analyzed in a phylogenetic tree. Interestingly, the various vertebrates ERV1-1 LTR sequences showed a very high nucleotide divergence when compared one to each other, preventing the construction of a NJ tree. The performed ML phylogenetic analysis (Figure 27) revealed that the vertebrates ERV1-1 LTRs are not related phylogenetically; showing no recognizable supported clusters (maximum bootstrap value = 33). This result is in line to the previous observation regarding the ERV-W group connection to the ERV1-1 sequences retrieved in *Platyrrhini* primates (Chapter 4): if, on the one side, a close phylogenetic relationship is shared between ERV1-1 and ERV-W retroviral internal portions (Figure 26); on the other side, the correspondent LTR sequences are characterized by an highly divergent structure when compared one to the others. This further suggests the possibility of a parallel but independent evolution of LTR elements as compared to the retroviral internal region.

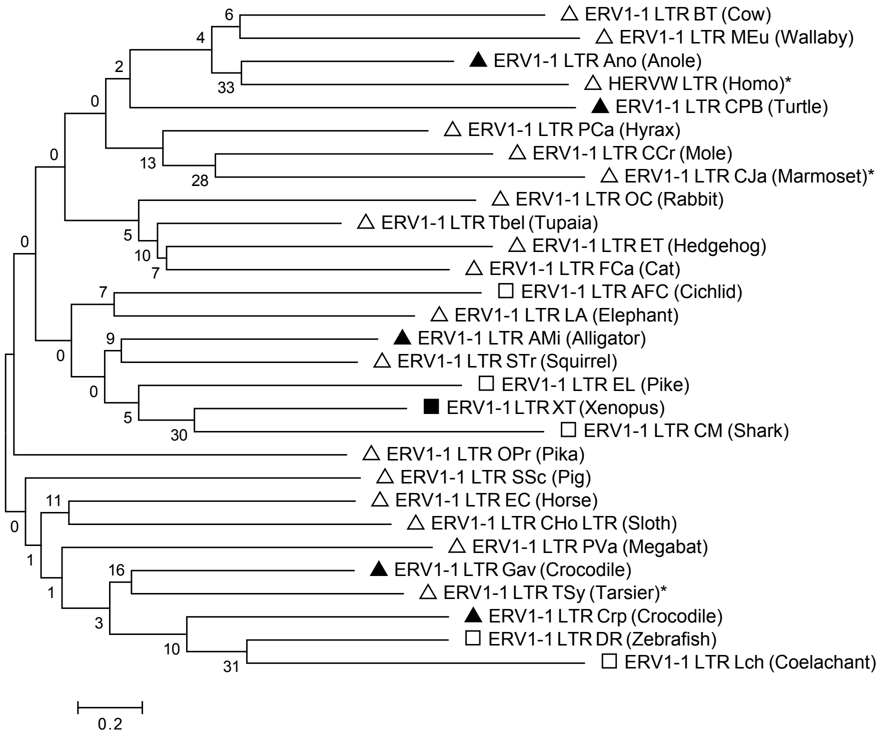

△ ERV1-1 LTR BT (Cow)
△ ERV1-1 LTR MEu (Wallaby)
▲ ERV1-1 LTR Ano (Anole)
△ HERVW LTR (Homo)*
▲ ERV1-1 LTR CPB (Turtle)
△ ERV1-1 LTR PCa (Hyrax)
△ ERV1-1 LTR CCr (Mole)
△ ERV1-1 LTR CJa (Marmoset)*
△ ERV1-1 LTR OC (Rabbit)
△ ERV1-1 LTR Tbel (Tupaia)
△ ERV1-1 LTR ET (Hedgehog)
△ ERV1-1 LTR FCa (Cat)
□ ERV1-1 LTR AFC (Cichlid)
△ ERV1-1 LTR LA (Elephant)
▲ ERV1-1 LTR AMi (Alligator)
△ ERV1-1 LTR STr (Squirrel)
□ ERV1-1 LTR EL (Pike)
■ ERV1-1 LTR XT (Xenopus)
□ ERV1-1 LTR CM (Shark)
△ ERV1-1 LTR OPr (Pika)
△ ERV1-1 LTR SSc (Pig)
△ ERV1-1 LTR EC (Horse)
△ ERV1-1 LTR CHo LTR (Sloth)
△ ERV1-1 LTR PVa (Megabat)
▲ ERV1-1 LTR Gav (Crocodile)
△ ERV1-1 LTR TSy (Tarsier)*
▲ ERV1-1 LTR Crp (Crocodile)
□ ERV1-1 LTR DR (Zebrafish)
□ ERV1-1 LTR Lch (Coelachant)
6
4
2
33
0
0
13
28
0
5
10
7
7
9
0
0
5
30
11
0
1
16
1
3
10
31
0.2

Figure 27. Phylogenetic relationships between the different RepBase ERV1-1 vertebrates groups (LTR portion). The consensus sequences representing the LTR portion of the ERV1-1 sequences retrieved in various vertebrate genomes were collected from RepBase Update. Among the vertebrates species taken into account, empty triangles indicate mammals, filled triangles indicate reptiles, empty squares indicate fishes and filled squares indicate amphibians. The RepBase entries corresponding to elements retrieved in primates genomes are further marked with an asterisk. Phylogenetic relationships were inferred on the full-length retroviral LTR portion nucleotide sequence by using the Maximum Likelihood method and the Kimura-2-parameter model. The resulting phylogeny was tested by using the Bootstrap method with 100 replicates. Length of branches indicates the number of substitutions per site

5.3 *Platyrrhini* ERV1-1 structural characterization

The close relation between the ERV1-1 groups in Marmoset and Squirrel Monkeys *Platyrrhini* and the ERV-W elements in *Catarrhini* primates genomes, observed at both nucleotide and amino acid level (Chapter 4, Figures 19 and 20), suggests a close evolutionarily connection between ERV1-1 and HERV-W Gammaretroviral groups.

In the light of this, we decided to characterize our ERV1-1 sequences dataset in more detail regarding the general retroviral structure, also considering that, at our best knowledge, no detailed information about this group are currently available in literature. The analyses were performed on Marmoset and Squirrel Monkey ERV1-1 most complete proviral sequences, i.e.

presenting complete LTRs, comprising 59 and 71 ERV1-1 elements, respectively. The ERV1-1 sequences showing truncated LTRs were also analyzed to assess the presence of processed pseudogenes, characterized by the already mentioned specific signatures within the LTRs and common in HERV-W group [33], but no such of these elements were found (data not shown).

The structural analysis of the Marmoset and Squirrel Monkey proviral ERV1-1 generated consensus (~ 9,3 Kb) led to the identification of the main retroviral regulatory and coding elements, whose nucleotide start and end positions are reported in Figure 28. The ERV1-1 group showed, as expected, a traditional retroviral structure, with *gag*, *pro*, *pol* and *env* genes flanked by the 5′ and 3′LTRs. The PBS, that during the retroviral infection binds a human tRNA to prime the reverse transcription process [296] is 18 nucleotides long and recognizes a R tRNA. The latter was already observed for an high proportion of HERV-W elements, being the second most frequent PBS type following the W one (Chapter 3 and ref. [21]).

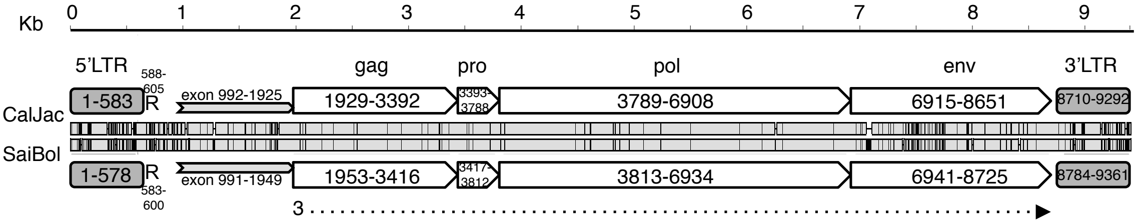


Figure 28. Main structural characteristics of ERV1-1 Marmoset (CalJac) and Squirrel Monkey (SaiBol) generated proviral consensus sequences.

LTRs, PBS and *gag, pro pol* and *env* genes nucleotide start and end positions are reported. Also the coordinates of a putative exon detected by RetroTector into the so-called *pre gag* portion is annotated. The reading frame for retroviral ORFs is indicated with a dotted arrow.

Beside the classical retroviral elements, both ERV1-1 consensus showed an untraditional region between 5′LTR-PBS and *gag* gene, termed here *pre gag* region, which is a common structural feature of HERV-W elements also [21]. Worth of note, within such *pre gag* region, ReTe analysis identified a putative exon at positions 992-1925 and 991-1949 in Marmoset and Squirrel Monkey consensus, respectively (Figure 28), and at positions 926-1838 in ERV1-1 RepBase reference sequence.

The search for the typical Gammaretroviral features in Marmoset and Squirrel Monkey ERV1-1 consensus highlighted the presence of i) one Gag NC Zinc finger with a $Cx_2Cx_4Hx_4C$ binding motif, usually involved in the retroviral RNA interaction during packaging [258], at positions 3219-3260 and 3243-3284, respectively; ii) a second Gag NC Zinc finger characterized by the loss of one of the variable residues ($Cx_2Cx_3Hx_4C$), as previously reported for both HERV-H [261] and HERV-W (Chapter 3 and ref. [21]) groups, at nucleotides 3291-3329 and 3315-3353, respectively; iii) a C-terminal Pol IN GPY/F motif (general structure: WxGPFxV), that binds the host DNA and could have a role in the integration specificity [259, 260], at positions 6685-6705 and 6711-6731, respectively; and iv) a nucleotide frequency bias, that could be determined by the action of encapsidated host RNA editing systems [257]. Particularly, ERV1-1 sequences showed a weak bias in purine amounts, with an enrichment in A (~ 28%) and a consequent impoverishment in G (~ 22%), as already reported for HERV-K [262] and HERV-W groups (Chapter 3 and ref. [21]).

5.3 ERV1-1 group phylogeny

In order to characterize the phylogenetic relations within the group, ERV1-1 proviral LTRs as well as *gag*, *pol* and *env* genes nucleotide sequences have been analyzed in a NJ tree with 1000 bootstrap replicates. As previously observed for HERV-W group (Chapter 3 and ref. [21]), while the retroviral genes analysis did not show statistically supported phylogenetic clusters, the 3′ and 5′ LTRs tree allowed us to further classify ERV1-1 group in at least 4 subgroups, named here A, B, C and D (Figure 29). In particular, subgroups A, B and C were well supported by bootstrap values of 99, 90 and 94, respectively. Subgroup B could be further divided into two clusters of elements, indicated in the tree as B1 and B2, with a 99 and 68 bootstrap support, respectively. Subgroup D showed instead a low statistical support, being also further divided in two clusters indicated as D1 and D2.

Beside these defined subgroups, a high proportion of ERV1-1 LTRs grouped into smaller clusters with principal nodes showing bootstrap values lower than 60. Interestingly, LTRs were often coupled with high bootstrap values (generally 99%), but the coupled LTRs were not necessary belonging to the same sequence (or even to the same primate genome), possibly suggesting events of recombination between different loci.

In order to better characterize the observed ERV1-1 LTR subgroups, we i) built a phylogenetic tree of the ERV1-1 subgroups members and the two HERV-W LTR subgroups (1 and 2) consensus (Chapter 3 and ref. [21]) using the NJ method and ii) compared each subgroup consensus with respect to the ERV1-1_CJa-LTR RepBase consensus sequence (Figure 30).

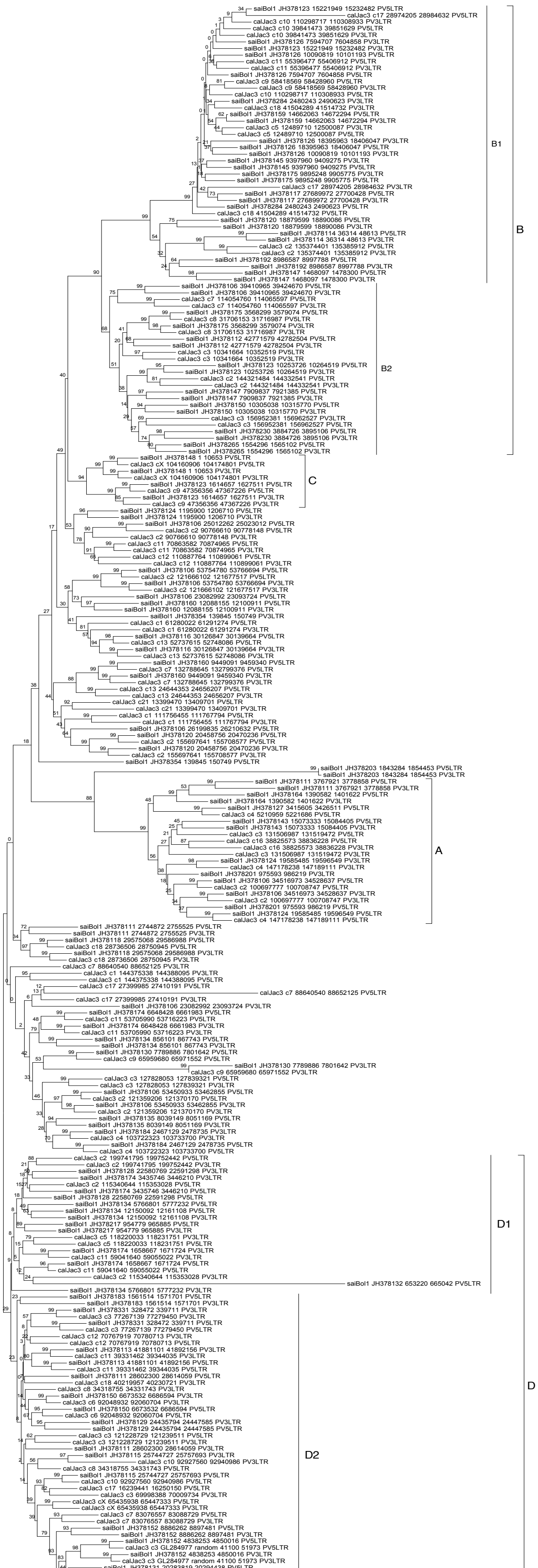


Figure 29. *Result of Marmoset and Squirrel Monkey ERV1-1 sequences LTRs phylogenetic analysis*

Marmoset and Squirrel Monkey ERV1-1 LTR nucleotide sequences phylogenetic relationships were inferred by using the Neighbor Joining method and the Kimura-2-parameter model and applying the pairwise deletion option. The resulting phylogeny was tested by using the Bootstrap method with 500 replicates. Length of branches indicates the number of substitutions per site

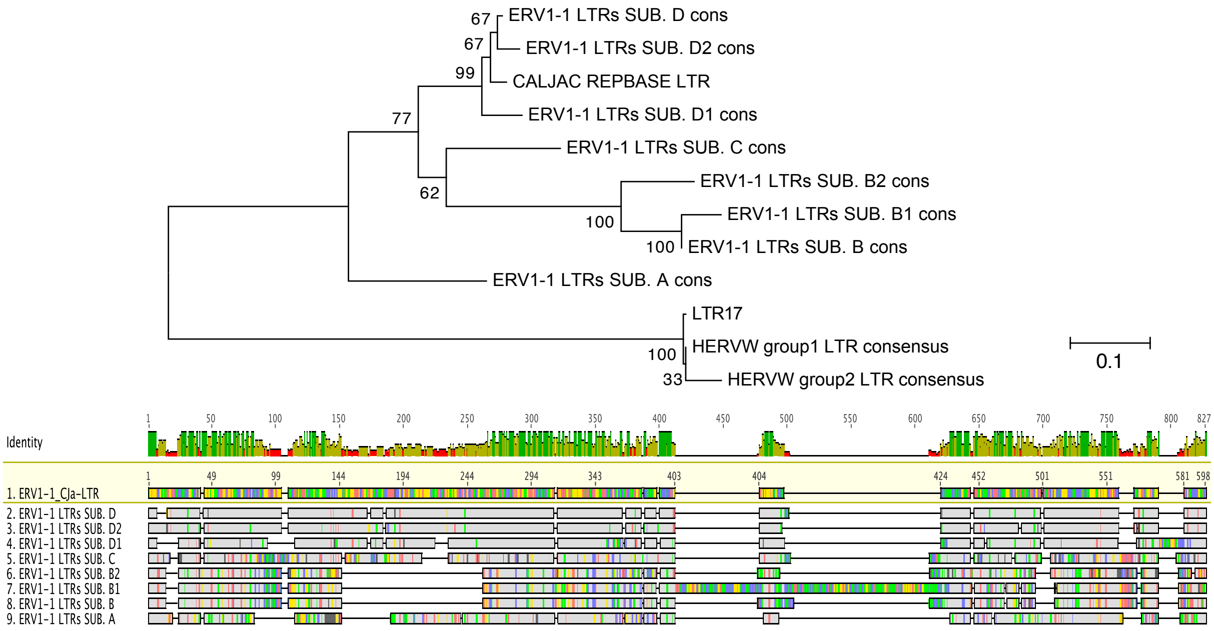


Figure 30. *Result of Marmoset and Squirrel Monkey ERV1-1 LTRs subgroups consensus phylogenetic analysis*
The nucleotide consensus sequences representing Marmoset and Squirrel Monkey identified ERV1-1 LTR subgroups phylogenetic relationships were inferred by using the Neighbor Joining method and the Kimura-2-parameter model and applying the pairwise deletion option. The resulting phylogeny was tested by using the Bootstrap method with 1000 replicates. Length of branches indicates the number of substitutions per site.

Based on our analysis, subgroup D elements are the most related to the RepBase ERV1-1_CJa-LTR, showing a lower number of substitutions as compared to the latter, also when the two clusters (D1 and D2) consensus sequences are considered. Subgroups B and C were phylogenetically related in the tree and, in fact, the consensus sequences alignment showed several nucleotide substitutions with respect to ERV1-1_CJa-LTR shared by both the relative consensus sequences, especially in the first 400 bases of the LTR structure. Interestingly, both subgroup B clusters (B1 and B2) sequences shared a recurrent deletion involving about 110 nucleotides, from base 146 to 255 of ERV1-1_CJa-LTR. In addition, B1 elements are characterized by a ~ 200 nucleotides insertion, found in the 80% of members. The analysis of such B1 elements insertion on Repeat Masking CENSOR tool [297] gave multiple results of homology with respect to different repeated elements, among which the most recurrent were Gypsy and HERVIP10. The presence of this insertion is also responsible for the two separate branches within the B1 cluster of sequences in the LTRs phylogenetic tree (Figure 29), where B1 insertion-devoid elements shared a lower degree of similarity (bootstrap support = 54) than the majority of B1 sequences that harbor the additional portion (bootstrap support = 99). As expected from the nucleotide identity plots (Chapter 4, Figure 19), HERV-W and ERV1-1 LTRs showed a highly dissimilar composition, differently with respect to the closely related internal retroviral portions. In Figure 30 tree, in fact, both the RepBase HERV-W group LTR consensus sequence (LTR17) and the two HERV-W subgroups LTR consensus sequences generated from our dataset (Chapter 3 and ref. [21]) clustered together in a separate, unrelated branch with respect to ERV1-1 subgroups LTRs consensus.

5.4 ERV1-1 proviruses estimated time of integration

The time of integration of the 81 collected ERV1-1 proviruses which clustered in the above mentioned subgroups was estimated based on the nucleotide divergence calculation, as described in Chapter 2 and 3. Briefly, for each ERV1-1 provirus, the percentage of nucleotide divergence (D) was calculated between i) the two LTRs of the same sequence, that are known to be identical at time of integration [254]; ii) each LTR and a generated LTR consensus for each subgroup of belonging, and iii) between the *gag* gene and a generated *gag* consensus for each subgroup of belonging. Regarding the two consensus-based approaches, given that the substitution rate acts randomly along each sequence, the subgroup-generated consensus should ideally represent the ancestral situation. To estimate each provirus time of integration (T), we divided the D values obtained from the three approaches by the *Platyrrhini* genomic substitution rate (SR), based on the relation T= D/SR. In addition, regarding the D between the 5′ and 3′LTRs of the same sequence, the obtained T value was further divided by a factor of 2, considering that each LTR evolved and accumulated mutations independently. In literature, a univocal SR for *Platyrrhini* genomic coding regions is currently not available, and no SR for *Platyrrhini* genomic non-coding regions has been defined yet. Thus, we decided to perform the initial age estimation considering i) two of the proposed *Platyrrhini* genomic coding regions reported in literature (0,14% and 0,126%) [221, 222], ii) a SR recently used to estimate HERV proviruses age (0,2%) [8] and iii) an averaged SR obtained from the previous three values (0,16%). The averaged value of the results obtained with the three age estimation approaches for each SR are graphically represented in Figure 31.

The use of different SRs led, obviously, to diverse ranges of estimated age for each subgroup. Given that the HERV-W/ERV1-1 sequences have been found in *Simiiformes* (*Platyrrhini* and *Catarrhini* parvorders) but not in the analyzed genomes of *Tarsiiformes* and *Prosimians*

(*Lemuriformes* and *Lorisiformes*) (Figure 17), we knew that these ERV groups time period of diffusion in primates lineage should be occurred before the evolutionarily divergence between *Platyrrhini* and *Catarrhini* parvorders, but after their separation with respect to *Tarsiiformes* and the most primitive *Prosimians*, i.e. approximately between 60 and 30/25 MYa. Hence, we supposed that SR values leading to an estimated time entry in primates lineage > 60 MYa and < 25 MYa could not be considered as reliable for this purpose. Based on this assumption, the most reliable SR for the ERV sequences time of integration estimation resulted 0,2% [8].

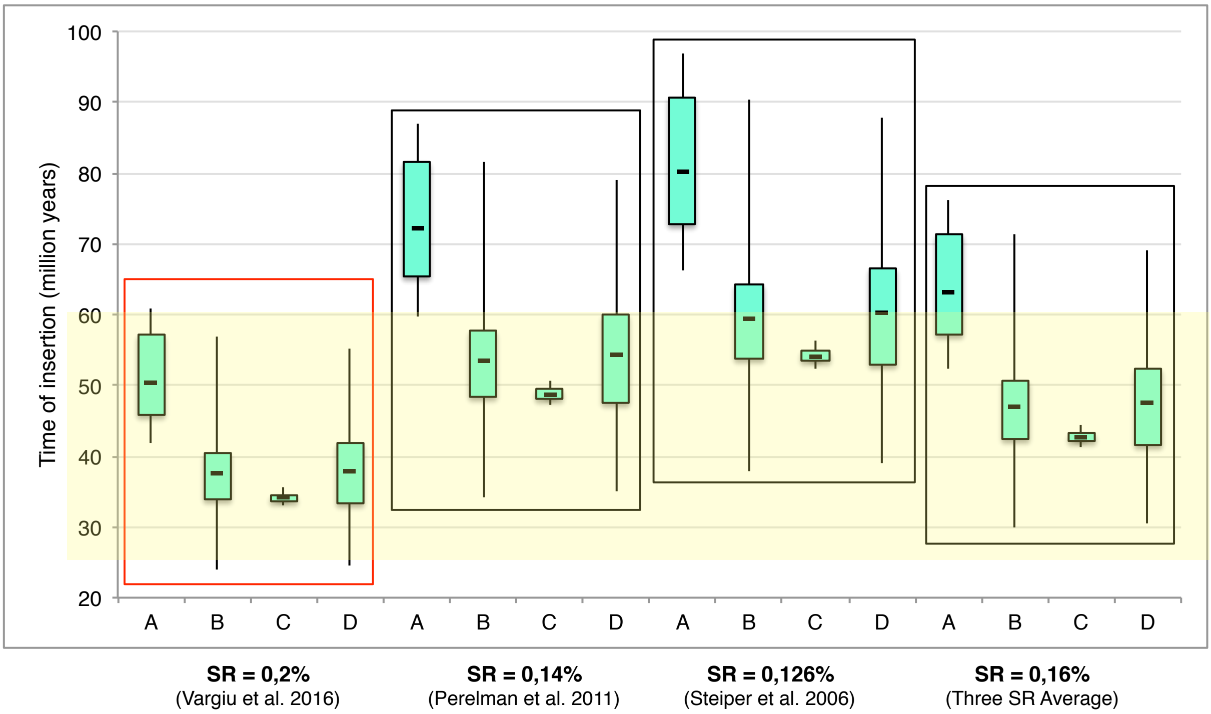


Figure 31. *ERV1-1 subgroups time of integration estimated based on different SRs for Platyrrhini primates genome.*
The group most probable period of diffusion, based on the presence of ERV1-1 sequences in *Platyrrhini* species but not in *Tarsiiformes* and *Prosimians* genome sequences, is highlighted in yellow. The SR values leading to estimated time of integration periods in disagreement with the observed presence of the group in evolutionarily related primates were not further considered. Thus, the most reliable SR for *Platyrrhini* Marmoset and Squirrel Monkey ERV sequences was 0,2% [8] (red line)

Results showed that ERV1-1 subgroup A proviruses constituted the first ERV1-1 insertions in Marmoset and Squirrel Monkey genomes, occurred mostly around 50-60 MYa and showing integration events during the following 10 MY. Differently, the other ERV1-1 subgroups were overall significantly younger (p=0,000000012) with respect to subgroup A members. In particular, subgroups B and D had median estimated time of integration around 38 MYa and an overall period of diffusion ranging between about 40 and 35 MYa; while subgroup C, characterized by a very few members, had median estimated time of integration around 34 MYa and an overall period of diffusion limited to about 33-36 MYa.

5.5 Discussion

The BLAT searches performed in Marmoset and Squirrel Monkey genome assemblies did not show the presence of any "true" ERV-W sequence in these *Platyrrhini* species, leading, however, to the identification of a group of ERV elements not previously known as related to ERV-W. Such sequences were listed as ERV1-1_CJa in Genome Browser, based on the RepBase annotations ERV1-1_CJa-I for the internal portion and ERV1-1_CJa-LTR for LTR sequences, and we subsequently referred to them as ERV1-1. In particular, a total of 130 ERV1-1 elements showing intact LTRs were retrieved in Marmoset and Squirrel Monkey genomes, and analyzed in terms of structure, phylogeny and estimated time of integration.

Noteworthy, all the retrieved ERV1-1 sequences were proviruses, and there was no evidence of L1-mediated processed pseudogenes formation, a mechanism that indeed massively interested the HERV-W group, accounting for around 2/3 of the latter insertions into the human genome [21, 33, 34]. Our previous analysis about the ERV-W sequences acquisition by the primates lineage revealed a quite extended period of activity of L1 elements, also in non-human primates genomes, with several species-specific insertions derived from new processed pseudogenes mobilization events. It is thus unlikely that such L1 machinery was instead not active in *Platyrrhini* species genome during the same time period. Hence, it is possible to speculate that the absence of ERV1-1 processed pseudogenes could be linked to the ERV1-1 nucleotide sequence composition, possibly presenting differences in the (still unknown) sites determinant for the recruitment and mobilization by L1 machinery. Considering the high nucleotide identity between the ERV1-1 and ERV-W proviral sequences along the traditional retroviral internal portion, such regions relevant for L1-mediated retrotransposition could eventually be located within the ERV-W LTRs and/or the *pre gag* regions, both presenting a rather divergent structure with respect to the other, still related, Gammaretroviruses. Hence, further analysis will be needed to clarify the specific determinants that made the massive L1 retrotransposition of ERV transcripts limited to the ERV-W group.

In the previous chapter, the high sequence similarity between the ERV-W and ERV1-1 groups and their close phylogenetic relation at both nucleotides and amino acids level suggested the possibility that these groups could possibly derive from a common ancestor. This possibility is further supported by the analysis of the ERV1-1 typical structural features, which resulted in common with the ones already characterized in the HERV-W group [21]. In particular, also in the majority of the ERV1-1 elements retrieved from Marmoset and Squirrel Monkey genomes, we found the presence of: i) two Gag NC Zinc fingers (one presenting a modified structure, as previously reported also in HERV-H [261] Gammaretrovirus), ii) one Pol IN GPY-F motif and a iii) a weak bias in purine amount [21]. Moreover, ERV1-1 sequences showed a PBS sequence generally recognizing an R tRNA, as frequently observed also for HERV-W group [21], which is, in fact, just slightly different from the W one [8]. Considering the taxonomic value of the above mentioned structural features, helpful to better understand retroviral phylogenetic relationships [257], their characterization in the ERV1-1 sequences could further indicate the presence of a common ancestor between this group and the ERV-W one.

The LTR-based phylogenetic analysis of Marmoset and Squirrel Monkey ERV1-1 proviruses showed the presence of at least 4 subgroups, named here with the letters from A to D. As compared to the RepBase consensus, Subgroup A members had the most divergent nucleotide sequence with respect to ERV1-1_CJa-LTR, which is indeed highly similar to subgroup D elements. Subgroups B and C sequences were more related one to each other, sharing several nucleotide substitutions as compared to ERV1-1_CJa-LTR. Interestingly, beside the above-mentioned subgroups, several LTRs were coupled with high bootstrap values (around 99%) even if they belonged to different loci, possibly reflecting previous events of LTRs recombination, or were even retrieved from the genome assembly of different *Platyrrhini* species.

After the ERV1-1 members phylogenetic classification, the time of integration of each ERV1-1 sequence was estimated through 3 approaches of divergence calculation: a traditional one, based on the divergence between the two LTRs of the same provirus, and two consensus-based methods considering LTRs and *gag* nucleotide sequences. Beside the characterization of the ERV1-1 group and the relative subgroups time period of acquisition, such analysis allowed also to identify the value of 0,2%, already used to estimate the age of various HERV groups [8], as most reliable substitution rate also for *Platyrrhini* primates ERV insertions. Results showed that subgroup A constitutes the older subset of ERV1-1 insertions, while the other elements were acquired significantly later on, being on average >10 MY younger. Overall, the whole group diffusion in *Platyrrhini* took place approximately between 55 and

35 MYa, with a first wave of insertions interesting subgroup A sequences approximately between 55 and 45 MYa, followed by the major wave of integrations that regarded subgroup B, C and D members along a time period ranging approximately from 40 to 35 MYa.

Overall, the analysis allowed to characterized for the first time ERV1-1 group structure and diffusion in the *Platyrrhini* species with available assembled genome sequences, further detailing their close evolutionarily relationship with the ERV-W group.

Chapter 6. Characterization of the HERV-W group pre gag element with respect to ERV1-1 and other Gammaretroviruses

6.1 Introduction

The evidence of a strong structural and phylogenetic relation between ERV1-1 and ERV-W groups led us to the structural characterization of ERV1-1 sequences, showing several retroviral features in common between the two groups. In particular, it is interesting to note that both sequences have an untraditional region located between the PBS and the *gag* gene, referred by us as *pre gag* portion during the HERV-W group characterization [21]. At our best knowledge, similarly to the HERV-W *pre gag*, no information is available about the ERV1-1 *pre gag* portion. In fact, for both elements, neither origin nor function has been proposed yet. Nevertheless, the *pre gag* portion is a stable component of both HERV-W and ERV1-1 groups, being found in almost all members and showing a considerable length (around 1,5-2 Kb), comparable to the one of retroviral coding genes. Indeed, it appears rather unlikely that such a big retroviral portion has been stably maintained without a reason, especially while the traditional retroviral genes were generally affected by huge recurrent deletions [21]. These considerations prompted us to further investigate the *pre gag* composition in these groups, and to search for a similar region in the other Gammaretroviral ERVs, in order to have more insights about such retroviral portion origin and possible ancestral function.

6.2 The *pre gag* region of HERV-W and ERV1-1 sequences

As previously mentioned, the HERV-W group is characterized by a ~ 2 Kb *pre gag* region, located between the PBS and *gag* gene [21] and being a stable component of the classical ERV-W structure, observed in almost all elements found in humans and *Catarrhini* primates. A similar element was also found to be a stable component of Marmoset and Squirrel Monkey ERV1-1 elements. In order to characterize these portions, we aligned and compared the HERV-W and ERV1-1 *pre gag* regions. Interestingly, differently form the close identity observed for the "traditional" retroviral internal portion, the HERV-W and ERV1-1 *pre gag* nucleotide identity is limited to approximately the first 400 bases (Figure 32).

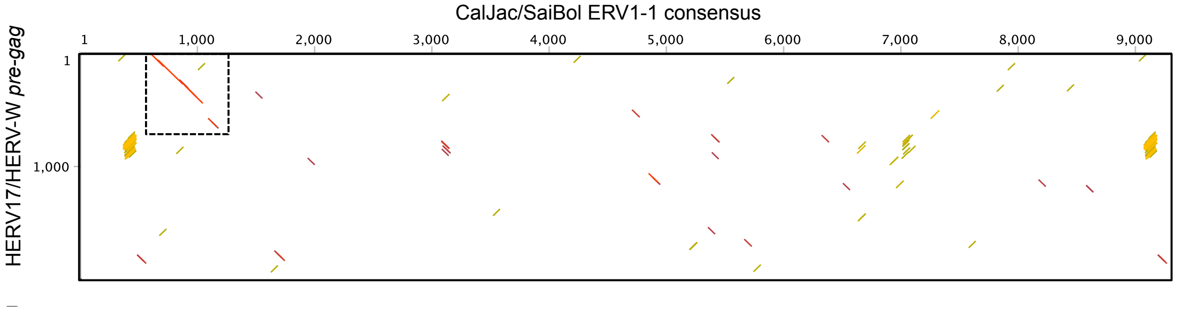


Figure 32. *Plots of nucleotide sequence comparison between HERV-W (HERV17) and Callithrix jaccus and Saimiri boliviensis ERV1-1 pre gag portions.*
Red lines represent the regions of nucleotide identity.

Worth of note, the ReTe analysis of HERV-W and Marmoset and Squirrel Monkey ERV1-1 proviral consensus sequences detected, in all of them, a putative exon co-localized with the *pre gag* region (Figure 33). In particular, in both ERV1-1 consensus sequences this putative exon spanned the *pre gag* region, comprising almost its whole length (from nucleotides 992 to 1925 and from nucleotides 991 to 1949, in ERV1-1 Marmoset and Squirrel Monkey consensus sequences, respectively). In the case of HERV-W consensus, the putative exon started into the *pre gag* portion and continued along the *gag* and *pro* genes, from nucleotide 1927 to 4305.

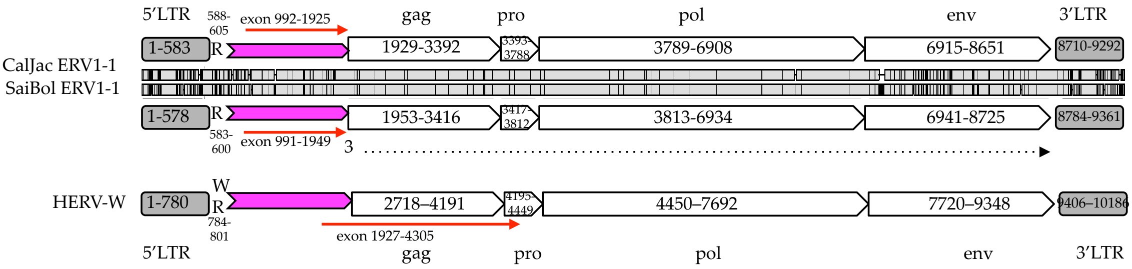


Figure 33*. Pre gag putative exon identified by ReTe analysis in ERV1-1 and HERV-W consensus sequences.*
For each sequence, the *pre gag* region is colored in pink and the putative exon is depicted as a red arrow. The start and end position of each structural element are also reported.

Importantly, due to the fact that the ERV1-1 and HERV-W *pre gag* regions share significant nucleotide identity only in approximately the first 400 bases (Figure 32), the putative ERV1-1 exon above described, being in the subsequent portion, was not found neither in human nor in other *Catarrhini* primates HERV-W *pre gag* regions, nor in any other genomic element (data not shown). Similarly, a Blast search of such exon predicted protein sequence, as translated in all 3 possible frames, gave no significant matches in *Catarrhini* genome sequences (data not shown).

6.3 Analysis of the HERV-W *pre gag* portion not shared with ERV1-1

We then wanted to assess the origin of the remaining HERV-W ~1.6 kb *pre gag* region with no homology with ERV1-1 *pre gag* sequence, and performed a Blat search using this sequence as a query in *Platyrrhini* Marmoset and Squirrel Monkey primates. Results showed that a portion of ~ 650 nucleotides in the HERV-W *pre gag* region has high score matches with another group of Squirrel Monkeys ERV sequences. Analysis of these sequences with RepeatMasker associated them to the human ERV group HERVIP10. Interestingly, this HERV-W *pre gag* 650 nucleotide region shared an overall 82% identity with the central portion of HERVIP10F ORF2 (approximate positions, nucleotides 3669-5160 in our generated SaiBol HERVIP10-like proviral consensus and nucleotides 3305-4768 in RepBase LTR10F-HERVIP10F-LTR10F consensus) (Figure 34). In particular, the HERVIP10F ORF2 encodes for a Pol-like protein, and the sequence shared with HERV-W *pre gag* corresponds to the 5′ portion of the RNase H domain based on ReTe analysis and GeneBank search for conserved motifs.

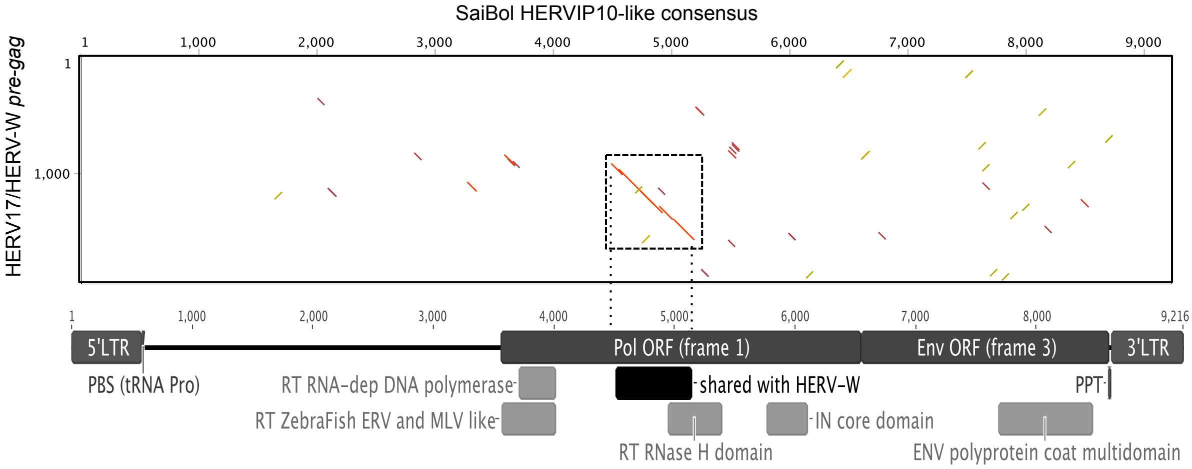
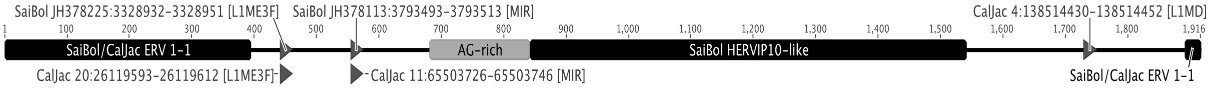


Figure 34. *Plots of nucleotide sequence comparison between HERV-W (HERV17) pre gag portion and Saimiri boliviensis HERVIP10 ORF2.*
Red lines represent the regions of nucleotide identity.

Of further note, while the RepeatMasker analysis of these HERVIP10-like SaiBol elements associated them to the HERVIP10 group if considering matches within the human genome, the same analysis performed considering *Platyrrhini* genomic sequences associated them not to HERVIP10 group, as expected, but to the RepBase Harlequin internal sequence. Harlequin elements are mosaic ERV sequences formed by portions with similarity to various HERV groups, comprising HERV-W, HERV-E, HERV-I and MER. We then asked whether the HERVIP10 or the HERV-W group members have provided the Harlequin HERVIP10-like portion to these mosaic recombinant sequences. The pairwise comparison of Harlequin HERVIP10-like region nucleotide sequence with both HERVIP10 and HERV-W consensus revealed the presence of a portion shared with the HERV-W consensus but not present within the RepBase HERVIP10 sequence, implying that the Harlequin HERVIP10-like region is probably derived from HERV-W sequences contribution. In light of this, we suggest that *Platyrrhini* HERVIP10-like sequences provided a non-ERV1-1 portion of the HERV-W *pre gag* sequence, and the latter contributed in turn in the Harlequin mosaic elements formation. Of note, despite in the RepBase Update description the HERVIP10 elements are reported to have been active about 30 MYa, their presence in Marmoset and Squirrel Monkey demonstrates an earlier diffusion, in line with their role of contributors for the HERV-W HERVIP10-like *pre gag* portion.

Taken together, the *Platyrrhini* ERV1-1 and HERVIP10-like elements contribution to the HERV-W *pre gag* region clarified the origin of about the 60% of the latter. Hence, also excluding another region rich in AG-repeats, constituting itself about the 8-9% of HERV-W *pre gag*, > 30% of the HERV-W *pre gag* sequence derived from a further contributor (Figure 35). The BLAT search of this remaining 30% of the HERV-W *pre gag* nucleotide sequence in *Platyrrhini* genomes showed just very short (20-30 nucleotides) stretches of similarity, generally related to other REs such as LINEs and MIR, but the subsequent RepeatMasker analysis did not confirm the presence of any RE in this *pre gag* portion (data not shown).


SaiBol JH378225:3328932-3328951 [L1ME3F]
SaiBol JH378113:3793493-3793513 [MIR]
CalJac 4:138514430-138514452 [L1MD]
1
100
200
300
400
500
600
700
800
900
1,000
1,100
1,200
1,300
1,400
1,500
1,600
1,700
1,800
1,916
SaiBol/CalJac ERV 1-1
AG-rich
SaiBol HERVIP10-like
CalJac 20:26119593-26119612 [L1ME3F]
CalJac 11:65503726-65503746 [MIR]
SaiBol/CalJac ERV 1-1

Element	start-end	contribution in LTR17-HERV17-LTR17[b]	contribution in HERVW[a]	pairwise similarity	*pre-gag* coverage
SaiBol ERV 1-1[a] CalJac ERV 1-1[a]	718-1125, 1914-1938 633-1038, 1890-1914	802-1196, 2693, 2717	803-1199, 2743-2767	60%	22%
AG repeats	/	1483-1643	1498-1691	/	9%
SaiBol HERVIP10-like[a] LTR10F-HERVIP10F-LTR10F[b]	3669-5160 3305-4786	1644-2343	1692-2396	82%	37%

Figure 35. *Summary of the HERV-W pre gag region composition.*
ERV1-1 and HERVIP-10 like portions are depicted as black blocks, while the AG-rich dinucleotides expansion as a light grey block. Dark grey triangles indicate aspecific short stretches of similarity with other repetitive elements, not confirmed by subsequent RepeatMasker analysis.

An exception was the ERV-W locus on Chimpanzee Y chromosome (positions 21951590-21956101), whose ERV1-1-like *pre gag* portion is extended of about 350 nucleotides, spanning in the ERV1-1 exon identified by ReTe. This ERV-W integration is annotated as LTR12F-HERV17-LTR12F based on RepBase; however, interestingly, its HERV17 internal portion presents the *pre gag* sequence derived from ERV1-1, but misses the following region with the AG-rich repeat and the HERVIP10-like portion, resulting to be actually more similar to ERV1-1 than to HERV17. In addition, differently from the HERV-W ones, this provirus LTRs, annotated as LTR12F, showed high nucleotide homology with ERV1-1 LTRs. Comparative genomic analysis localized a corresponding orthologous sequence in human locus Yq11.221, at positions 16452374-16456884, being also annotated as LTR12F-HERV17-LTR12F. This locus and other elements with similar structure were included within the Table 17 Blat-retrieved elements showing low score identity to HERV17, and could possibly represent a subset of evolutionarily intermediate elements with a structure closer to the ERV1-1 sequences one. This could also explain their missed detection by the previously performed Blat search in the human genome sequence using the HERV-W group consensus sequences as queries.

Finally, it is interesting to note that a minority of HERV-W loci, entirely constituted by processed pseudogenes, lacks the whole *pre gag* region (Figure 36). The absence of this portion was also confirmed in the correspondent non-human *Catarrhini* primates loci, which genomic positions are reported in Table 6. The fact that all the (H)ERV-W loci lacking the *pre gag* portion are actually processed pseudogenes possibly suggests that the *pre gag* portion have been removed by splicing mechanisms from the transcripts originating such elements, being processed as an intronic region. However, the fact that the great majority of processed pseudogenes present instead such element also indicate the existence of alternative splicing variants retaining the entire *pre gag* region, commonly to exonic portions.

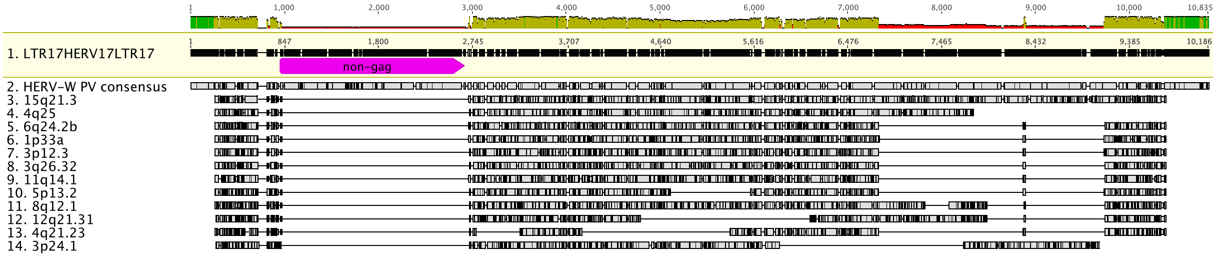


Figure 36. *HERV-W processed pseudogenes avoid of the entire pre gag region.*
Sequences are aligned to LTR17-HERV17-LTR17 consensus sequence

6.4 Presence of the *pre gag* region in other Gammaretroviral HERVs

Beside the HERV-W group [21], the presence of a *pre gag* portion was previously reported for HERV-H Gammaretroviruses [261]. We then asked whether such region could be a common feature of all Gammaretroviral HERVs, possibly suggesting a functional role of *pre gag* elements in the relative ancestral exogenous viruses.

Hence, the proviral consensus sequences obtained during HERV-W and ERV1-1 group characterization were aligned with respect to the RepBase reference sequences of various endogenous human Gammaretroviruses (HERV-W, HERV9, HERV30, HERV-H, HERV1, HERV3, HERV-L, HERV-E, Harlequin, HERV-T, HERV-H48, MER50, MER57, PRIMA41, PABL, HERVP71A, HERV-Fc1, HERV-FH21) and GenBank Baboon ERV consensus sequence (BaEV, gi|61506|emb|X05470.1|).

As shown in Figure 37, the *pre gag* portion shared between ERV1-1 and HERV-W sequences was present also in HERV9 and HERV30 reference sequences, confirming their close relation to HERV-W and ERV1-1 groups as shown in previous phylogenetic analysis. Differently, all the other Gammaretroviruses *pre gag* portions showed a heterogeneous nucleotide composition, presenting low identity one to each other and with respect to ERV1-1 and HERV-W *pre gag* elements.

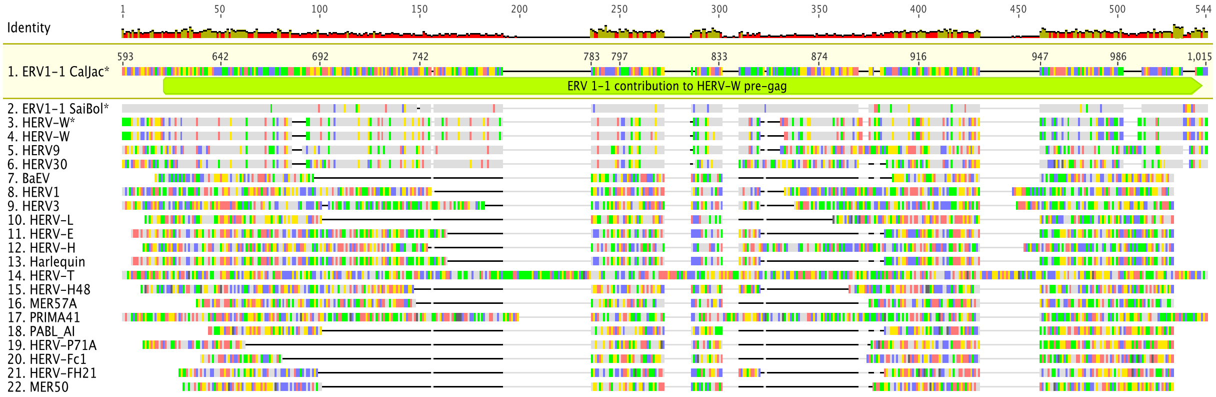


Figure 39. *Comparison of ERV1-1 Marmoset (CalJac) and Squirrel Monkey (SaiBol) pre gag portion with the pre gag region of HERV-W and the other Gammaretroviruses.*

The consensus sequences marked with a * were generated from the proviral datasets used in this study; the other HERV consensus sequences were retrieved from RepBase Update database except for Baboon HERV sequence (BaEV) that was retrieved from GenBank.

Nucleotide identity between the consensus sequences is represented by the colored upper bar (green: 100% identity; greeny-brown: between 100% and 30% identity; red: identity < 30%), divergent nucleotides are indicated by colored lines.

6.5 Discussion

During the characterization of the HERV-W group within the human genome, we observed the stable presence of an untraditional so-called *pre gag* region, comprised between the PBS sequence and the *gag* gene and found in almost all HERV-W sequences (being either proviruses or processed pseudogenes). The presence of such HERV-W ~2 Kb *pre gag* region was originally reported for three cDNA HERV-W clones [23], but neither function or origin were never proposed. Moreover, a similar region, at least regarding its localization, was previously observed for another Gammaretrovirus-like HERV group, namely HERV-H [261]. Also at the light of the identification of a similar *pre gag* element in ERV1-1 proviruses (Chapter 5, Figure 28), we decided to investigate in further detail the HERV-W *pre gag* region and to compare it with the ERV1-1 and the HERV-H one as well as with eventual homologous portions in various Gammaretroviral HERVs.

We hence report that, contrarily to the traditional HERV-W retroviral internal portion, the HEVR-W *pre gag* region shares close nucleotide identity to the ERV1-1 *pre gag* only along the ~400 nucleotides. In addition, in both ERV1-1 and HERV-W *pre gag* regions ReTe identified a putative exon, whose exons bioinformatically-translated protein product showed no similarity with any known protein, neither of human nor of retroviral origin. The ERV1-1 *pre gag* exons were localized within the respective Marmoset and Squirrel Monkey *pre gag* portion lacking nucleotide homology with respect to HERV-W, and were thus generally not found in *Catarrhini* (H)ERV-W sequences. Nevertheless, an (H)ERV-W locus found within Chimpanzee and human Y chromosome presented a portion of the ERV1-1 *pre gag* putative exon, and the pairwise nucleotide comparison of its whole sequence (internal portion and LTRs, annotated as HERV17 and LTR12 respectively by RepeatMasker) revealed higher similarity to ERV1-1 than to ERV-W. The same locus was included in the group of HERV-W-like sequences identified in humans by comparative genomics localization of non-human *Catarrhini* Blat-retrieved sequences showing low score identity to HERV17 (Chapter 4, Table 17). All these evidences, together with the fact that LTR divergence calculation leads to an estimated integration time >60 MYa, suggest that this locus could represent a more "ancestral" form among ERV-W members and that, possibly due to its integration in a low-rate recombination chromosome, had maintained characteristics typical of the group progenitor, being more similar to ERV1-1 instead of HERV-W.

Beside the above mentioned HERV-W *pre gag* initial portion, sharing high identity to the first ~400 bp of the ERV1-1 *pre gag*, a further ~650 bp portion of the HERV-W *pre gag* region was found as highly similar to another *Platyrrhini* ERV group. Particularly, HERVIP10F ORF2, which encodes for a Pol-like protein, shares a 82% nucleotide similarity with the HERV-W *pre gag* in the region encoding for the RH domain. Of note, it has been reported that LTR-retrotransposons (including HERVs) originally harbored a non-LTR-retrotransposon-derived RNase H domain [298]. During evolution, the lineage of elements leading to vertebrate retroviruses acquired a new RNase H domain, either from non-LTR retrotransposons or from a eukaryotic host genome, and the preexisting RNase H domain degenerated to become the RT tether domain [298]. It is hence possible to speculate that a similar, also if currently unexplainable, event could led to the presence of the HERIP10 RH portion in HERV-W *pre gag*. Despite the fact that the HERVIP10F group activity has been reported as occurred about 30 MYa based on RepBase information, the presence of HERVIP10-like elements in Squirrel Monkey genome suggested that the group had instead an earlier diffusion, in accordance with its contribution to the HERV-W *pre gag* structure. Interestingly, the HERVIP10-like *pre gag* portion is absent in the ERV1-1 structure, indicating that its acquisition occurred in *Catarrhini* ERV-W elements only. The further mobilization of the HERVIP10-derived *pre gag* portion from HERV-W structure is then suggested by its presence as a component of Harlequin mosaic elements. A similar recombination event, involving a still unknown LTR donor, could have provided the ERV-W LTRs in *Catarrhini* primates.

If excluding the ERV1-1- and HERVIP10-like portions in the HERV-W *pre gag* region, and without considering the variable expanded simple repeats regions, about the 30% of the HERV-W *pre gag* derives from a further, still not identified third contributor.

Remarkably, the multiple alignment of various Gammaretroviral HERV group consensus sequences showed the presence of a *pre gag* portion in other groups than HERV-H, HERV-W and ERV1-1, possibly indicating the *pre gag* component as a characteristic feature shared by the ancestral exogenous Gammaretroviruses. The nucleotide sequence comparison of these various Gammaretroviral consensus *pre gag* regions showed however an heterogeneous composition, but the ERV1-1 *pre gag* portion shared with HERV-W presented high identity also to the correspondent *pre gag* regions of HERV9 and HERV30. This further suggest that these Gammaretroviral HERVs could share a common ancestor, being strongly related at both nucleotide and protein level.

In conclusion, the maintenance of such a big region as a stable component of various Gammaretroviral HERVs, together with the presence of a putative exon with unknown function observed in the HERV-W and ERV1-1 *pre gag* portions, could indicate an original important role of the *pre gag* elements in the ancestral Gammaretroviruses biology, which needs be further investigated.

Chapter 7. Conclusions

Since the discovery of the HERV-W specific expression in placenta and, especially, given the HERV-W functional Env (Syncytin-1) exapted during evolution for the syncytiotrophoblast formation, a great attention has been devoted to the HERV-W group expression at both RNA and protein level. Many independent studies assessed the overall group expression among a wide amount of human healthy as well as pathological tissues, through the use of different methodologies. Particularly, the great majority of these studies were aimed to demonstrate the group general expression in diseased tissues, primarily to find a connection with a number of human pathologies, including cancers, autoimmune and infectious diseases. As an example, more than 40 studies were performed to tentatively assess the HERV-W expression role in MS, as recently ad exhaustively reviewed [299]. Despite this great attention and the many efforts done, neither HERV-W nor any other HERV groups have been reliably connected to any human disorder yet. As commonly observed also for the other HERV groups, a major obstacle in the evaluation of HERV-W group expression was the lack of an exhaustive characterization of the single loci integrated into the human genome. The available information about the HERV-W group presence at the genomic level, in fact, were still referred to analyses performed a number of years ago, using unstandardized methodologies and providing partial results very difficult to compare with current data. This prevented, until now, the association of the observed expression profiles to the precise HERV-W locus of origin, essential for the definitive demonstration of an eventual specific role in the various physio-pathological contexts.

At the light of this, we decided to analyze in great detail the HERV-W group single members sequence, characterizing each of them in terms of structure, context of insertion, phylogeny and estimated time of entry into the human genome.

A total of 213 HERV-W sequences were retrieved from human genome assembly GRCh37/hg19 and each single nucleotide insertion/deletion was characterized, providing a unique dataset that allows now to unambiguously identify each HERV-W integration (and its expressed products). The HERV-W sequences structural characterization allowed also to describe, for the first time, the group typical features in a complete dataset and in an updated version of the human genome. Interestingly, the presence of a second Gag NC Zn-finger with a modified amino acid motif was not reported previously for the HERV-W group. Moreover, the structural analysis highlighted a number of recurrent deletions affecting the retroviral genes, and a series of recurrent insertion/deletions and nucleotide substitution within a subset of LTR sequences. In addition, the LTRs phylogenetic analysis allowed to identify two HERV-W phylogenetic subgroups (1 and 2), showing that subgroup 2 LTRs can be recognized by the presence of the above mentioned key mutations, having a value as phylogenetic features. Remarkably, the group period of diffusion has been characterized in detail through the estimation of the integration time of each single HERV-W sequence. The use of a multiple approach of divergence calculation allowed the inclusion of the 94% of HERV-W elements (instead of the 23% if considering only the traditional 5′ and 3′ proviral LTRs divergence based method), showing a rather extended time period of diffusion in primates, as occurred between 43 and 30 MYa with sporadic insertion until <10 MYa. Particularly, the first events of HERV-W integration involved subgroup 2 sequences; with subgroup 1 elements being instead significantly younger ($p<0,0005$) with a colonization of the human genome occurred in average about 8 million years later. To investigate the single HERV-W members potential effects on the human genome, their context of insertion as well as their predicted binding to human TFs were characterized. A total of 80 HERV-W elements were found as localized within human genes, being integrated into 55 coding genes and 25 non-coding genes. With respect to the holding human gene, we observed both a strong antisense orientation and an intronic integration biases for the HERV-W sequences inserted into coding regions, resulted absent for the HERV-W loci inserted into non-coding genes. The latter were reported as co-localized with human non-coding genes for the first time in the great majority of cases, and can possibly influence the host biology even in the absence of an expressed product. At least 16 HERV-W sequences LTRs were predicted as able to bind various human TFs at high score values, 8 of them being also localized within human genes. Finally, all the 213 HERV-W sequences as well as the publically available MSRV cDNA clones and probes were analyzed regarding the completeness of their *env* ORFs and the predicted puteins. With the exception of Syncytin-1, all the analyzed ORFs harbored internal stop codons and/or reading frameshifts, and the relative puteins were affected by various substitutions, particularly frequent in regions with known functional activity in placenta. Furthermore, the previously reported MSRV clones and probes were compared to our complete dataset, identifying the possible locus of origin or, alternatively, the loci potentially involved in multiple transcripts recombination, a frequent complication to be taken into account in the general analysis of expressed multicopy elements.

Following the characterization at human genomic level, a further step of the present work was the comparative analysis of the HERV-W sequences dynamics and periods of diffusion in the genomes of evolutionarily related primates, belonging to *Simiiformes* (*Catarrhini* and *Platyrrhini* parvorders), *Tarsiiformes* and *Prosimians*. Starting from each human locus, the corresponding ERV-W orthologous insertions were localized in 5 non-human *Catarrhini* species, providing the first comprehensive comparative map of 211 ERV-W orthologous loci (no reliable comparative information was available for the two loci within Y chromosome). In addition, the HERV17 Blat search in each non-human *Catarrhini* species revealed various events of solitary LTRs formation and species-specific new acquisitions occurred during primates speciation, further detailing the group dynamics of diffusion. Interestingly, both the identification of each ERV-W orthologs primate species of first acquisition and the presence of various ERV-W species-specific insertions confirmed that the ERV-W colonization of primates lineage occurred within an extended period of time, having peaked between ~42 and 30 MYa and providing sporadic, species or lineage-specific ERV-W locus formations until less than 10 MYa. In this context, interestingly, the L1-mediated formation of processed pseudogenes was a major phenomenon occurred in parallel to new proviral integrations, accounting for >2/3 of both orthologous and species-specific ERV-W loci present within the primate lineages. This revealed that processed pseudogenes formation greatly contributed to the rather long proliferation activity of the ERV-W group instead of accounting for its lower proliferative success, contrarily to what previously thought.

The search for ERV-W sequences in *Platyrrhini* gave negative results, but allowed to identify a group, named ERV1-1 based on RepBase annotations, showing an unreported high nucleotide identity to the ERV-W internal retroviral portion. The presence of a close evolutionarily relationship between ERV1-1 and ERV-W groups was confirmed through the phylogenetic analysis of their retroviral predicted proteins (puteins) with respect to the other Gammaretroviruses, even if also HERV9 and HERV30 resulted highly related to both groups. The strong relationship between ERV1-1 and ERV-W elements was supported also by the characterization of 130 ERV1-1 insertions retrieved in Marmoset and Squirrel Monkey genome assemblies, leading to the first description of the ERV1-1 group structure, phylogenetic subgroups and estimated time of integration. Noteworthy, the analysis showed several structural features shared between the (H)ERV-W and ERV1-1 groups, among which a second Gag NC Zinc finger with modified structure, a Pol C-terminal IN GPY/F motif and a weak bias in purines amount. Moreover, as previously done for the HERV-W group, the

ERV1-1 LTRs sequence and phylogenetic analysis allowed to classify the ERV1-1 sequences into 4 subgroups. The time of integration of the single ERV1-1 members was estimated, showing that subgroup A had the most ancient integrations while subgroups B, C and D members were acquired in average >10 MY later on, with an overall diffusion of the whole group occurred approximately between 55 and 35 MYa.

Of further note, another untraditional retroviral element found in the great majority of both HERV-W and ERV1-1 sequences was the so-called *pre gag* region, located between the 5'LTR and the *gag* gene and for which neither origin nor function have been proposed yet. We analyzed HERV-W and ERV1-1 *pre gag* elements, showing that they share high nucleotide identity in the first ~400 bp portion. Another ~600 bp *pre gag* region was found as highly similar to another *Platyrrhini* ERV group, identified in Squirrel Monkey genome assembly RepBase annotations as Harlequin but being highly related to the HERVIP10 group based on our analysis. Interestingly, the portion sharing close nucleotide similarity to the HERV-W *pre gag* belonged to the HERVIP10 ORF2, encoding for a Pol-like protein. The identification of a putative exon in the *pre gag* elements of both ERV1-1 and HERV-W together with the presence of a similar, untraditional retroviral portion in many Gammaretroviral HERVs further suggests a possible important role of this element in the ancestral Gammaretroviruses biology, and, eventually, a residual potential biological significance in the human genome, deserving to be further investigated with functional studies.

Overall, the present work provides a detailed picture of ERV-W and the related ERV1-1 groups structure and diffusion in primates, with unreported insights about the ERV-W sequences origin, composition and evolution up to their presence in the human genome. The generated exhaustive and updated HERV-W dataset could be an important and reliable background to unambiguously identifying the uniqueness of each HERV-W integration, and to connect the observed expression products to the single loci of origin. This could be particularly useful to finally characterize the specific HERV-W loci expression in both physiological and pathological contexts, and to definitively assessing their contribution to the human transcriptome and the eventual mechanistic role in human diseases. Moreover, the precise description of the single HERV-W loci structural characteristics and context of integration could help in evaluating their effective expression potential and the possible consequences on the host biology beside the presence of any expressed products. This could be particularly valuable to definitively assess the HERV-W role in human physio-pathological contexts, and to identify specific sequences suitable as biomarkers or innovative therapeutic targets for a number of human disorders.

Moreover, it would be interesting to complete the analysis of the HERV-W group diversity and impact on the host by the characterization of the single HERV-W sequences polymorphic variants in the human population, in order to understand their possible influence on the inter-individual genetic variability and susceptibility to those diseases potentially connected to HERVs presence and expression. Finally, considering that many of these pathological contexts are influenced by an altered hypomethylated epigenetic environments, possibly liberating HERVs expression, also the specific dynamics of the L1-mediated retrotransposition of HERV-W transcripts should be investigated in detail, allowing to evaluate the eventual occurrence of de-novo insertional mutagenesis events. Importantly, bioinformatics constitutes a valuable and powerful tool for the rational characterization of multicopy elements held by the human genome, providing a reliable background for their subsequent specific analysis by the traditional wet-lab approaches.

The state of the art and the results presented in this work have been published/submitted in/to the following peer-reviewed international journals:

1. Grandi N, Tramontano E: **Type W Human Endogenous Retroviruses (HERV-W) proviral integrations and their mobilization by L1 machinery: contribution to the human transcriptome and impact on the host physio-pathology**. *Viruses* 2017 (invited review) [Submitted]

2. Grandi N, Cadeddu M, Blomberg J, Mayer J, Tramontano E: **HERV-W group evolutionary history in non-human primates: characterization of ERV-W orthologs in *Catarrhini* and related ERV groups in *Platyrrhini*.** *BMC Evolutionary Biology* 2017 [Submitted]

3. Grandi N, Cadeddu M, Blomberg J, Tramontano E: **Contribution of type W human endogenous retrovirus to the human genome: characterization of HERV-W proviral insertions and processed pseudogenes**. *Retrovirology* 2016, **13**:1–25.

And is included in two additional manuscripts currently in preparation:

4. Grandi N, Demurtas M, Blomberg J, Mayer J, Tramontano E: **Characterization of the ERV1-1 sequences in Marmoset and Squirrel Monkey (Platyrrhini parvorder): further insights into the group close phylogenetic relation to ERV-W.** [Manuscript in preparation]

5. Grandi N, Blomberg J, Mayer J, Tramontano E: **Characterization of the HERV-W group pre gag element with respect to ERV1-1 and other Gammaretroviruses.** [Manuscript in preparation]

Acknowledgements

I would like to thank Prof. Enzo Tramontano for being an excellent scientist and a brilliant mentor, who guided me and made me constantly grow during these three years.
Many thanks also to Prof. Jonas Blomberg and Prof. Jens Mayer for being such great collaborators and for having taught me so many things. Jens, thank you also for the hospitality and the support you gave me during my time period in Homburg.
Of course, I would like to thank all my colleagues and all the students that worked with us every day at the Laboratory of Molecular Virology. In particular, special thanks to Francesca, Angela, Maria Paola, Elisa and Martina for making every working day a funny day also.

I would like also to thank all the people who have always supported and encouraged me during this journey:

A million of thanks to Francesco, love and partner of my life, for being my family and for having made everything possible, bringing me to this amazing Island that now I call home and being at my side every single day. You supported and encouraged me during these three years, often sacrificing our time together to support me in this fantastic - but also so challenging - work, and making me happy and proud to be myself with your love and kindness.
Thanks to my wonderful parents, Sonia and Renzo, my beloved sister Celeste, my fantastic aunt Simonetta, my stupendous grandparents and all my family for making me what I am and for being proud of me everything I do.
Many thanks also to Igino and Gabriella, for being such wonderful people who made me feel at home treating me like a daughter since the very first day.
I love all of you so much.

Last but not least, I would like to thank all my new and old friends:

Thanks to all my friends in Bologna, that will have always a special place in my heart, and thanks to all the wonderful people that I had the opportunity to know here in Sardinia, who have welcomed me so kindly and made me immediately feel part of this amazing place.
Special thanks to Alice Elena and Andrea, the best friends one can desire, for being so wonderful and irreplaceable, always ready to take a plane only to surprise me. Even so far away, you are by my side every day, and have you here in this special day is the best gift you could get me. Love you so much guys!

Nicole

Ringraziamenti

Vorrei ringraziare il Prof. Enzo Tramontano per essere uno scienziato eccellente e un mentore brillante, che mi ha guidato e mi ha fatto crescere costantemente nel corso di questi tre anni.
Molti ringraziamenti anche al Prof. Jonas Blomberg e il Prof. Jens Mayer per essere collaboratori grandiosi e per avermi insegnato così tante cose. Jens, grazie anche per l'ospitalità e il sostegno durante il mio periodo alla Saarland University ad Homburg.
Naturalmente, vorrei ringraziare tutti i miei colleghi e tutti gli studenti che hanno lavorato con noi ogni giorno presso il Laboratorio di Virologia Molecolare. In particolare, un ringraziamento speciale va a Francesca, Angela, Maria Paola, Elisa e Martina per sapere rendere così divertente ogni giorno lavorativo.

Vorrei inoltre ringraziare tutte le persone che hanno mi sempre sostenuto ed incoraggiato durante questo viaggio:

Un milione di grazie a Francesco, amore e compagno della mia vita, per essere la mia famiglia e per avere reso tutto possibile, portandomi in questa straordinaria isola che ora io chiamo casa e rimanendo al mio fianco ogni giorno. Mi hai sostenuto ed incoraggiato nel corso di questi tre anni, spesso sacrificando il nostro tempo insieme per sostenermi in questo fantastico - ma anche tanto impegnativo - lavoro, e mi hai reso felice e orgogliosa di essere me stessa con il tuo amore e la tua bontà.

Grazie ai miei meravigliosi genitori, Sonia e Renzo, alla mia amata sorella Celeste, alla mia fantastica zia Simonetta, ai miei nonni stupendi e a tutta la mia famiglia per avermi resa quello che sono e per essere fieri di me qualsiasi cosa faccia.
Molte grazie anche a Igino e Gabriella, per essere persone tanto meravigliose e per avermi fatto sentire a casa trattandomi come una figlia fin dal primo giorno.
Amo moltissimo tutti voi.

Ultimo ma non meno importante, vorrei ringraziare tutti i miei vecchi e nuovi amici:

Grazie a tutti gli amici di Bologna, che avranno sempre un posto speciale nel mio cuore, e grazie a tutte le persone meravigliose che ho avuto la possibilità di conoscere qui in Sardegna, che mi hanno accolto così gentilmente facendomi sentire subito parte di questo luogo incredibile.

Un ringraziamento speciale ad Alice Elena e Andrea, i migliori amici che si possano desiderare, per essere così meravigliosi ed insostituibili, sempre pronti a prendere un aereo solo per farmi una sorpresa. Anche da lontano, siete al mio fianco tutti i giorni, e avervi qui in questo giorno speciale è il miglior regalo che poteste farmi. Vi voglio tanto bene ragazzi!

Nicole

Bibliography

1. International Human Genome Sequencing Consortium: **International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**:931–945.

2. Cordaux R, Batzer MA: **The impact of retrotransposons on human genome evolution.** *Nat Rev Genet* 2009, **10**:691–703.

3. Ohshima K: **RNA-Mediated Gene Duplication and Retroposons: Retrogenes, LINEs, SINEs, and Sequence Specificity.** *Int J Evol Biol* 2013, **2013**:424726.

4. Bannert N, Kurth R: **The Evolutionary Dynamics of Human Endogenous Retroviral Families**. *Annu Rev Genomics Hum Genet* 2006, **7**:149–73.

5. Mager DL, Goodchild NL: **Homologous recombination between the LTRs of a human retrovirus-like element causes a 5-kb deletion in two siblings.** *Am J Hum Genet* 1989, **45**:848–854.

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

7. Sperber G, Airola T, Jern P, Blomberg J: **Automated recognition of retroviral sequences in genomic data - RetroTector©**. *Nucleic Acids Res* 2007, **35**:4964–4976.

8. Vargiu L, Rodriguez-Tomé P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, Tramontano E, Blomberg J: **Classification and characterization of human endogenous retroviruses; mosaic forms are common.** *Retrovirology* 2016, **13**:7.

9. Magiorkinis G, Belshaw R, Katzourakis A: **"There and back again": revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era**. 2013.

10. Schön U, Seifarth W, Baust C, Hohenadl C, Erfle V, Leib-Mösch C: **Cell Type-Specific Expression and Promoter Activity of Human Endogenous Retroviral Long Terminal Repeats**. *Virology* 2001, **279**:280–291.

11. Stauffer Y, Theiler G, Sperisen P, Lebedev Y, Jongeneel CV: **Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues.** *Cancer Immun a J Acad Cancer Immunol* 2004, **4**:2.

12. Seifarth W, Frank O, Zeilfelder U: **Comprehensive analysis of human endogenous retrovirus transcriptional activity in human tissues with a retrovirus-specific microarray**. *J Virol* 2005, **79**:341–352.

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

14. Hu L: *Endogenous Retroviral RNA Expression in Humans*. 2007.

15. Pérot P, Mugnier N, Montgiraud C, Gimenez J, Jaillard M, Bonnaud B, Mallet F: **Microarray-based sketches of the HERV transcriptome landscape**. *PLoS One* 2012, **7**:haase.

16. Haase K, Mösch A, Frishman D: **Differential expression analysis of human endogenous retroviruses based on ENCODE RNA-seq data**. *BMC Med Genomics* 2015, **8**:71.

17. Balestrieri E, Pica F, Matteucci C, Zenobi R, Sorrentino R, Argaw-Denboba A, Cipriani C, Bucci I, Sinibaldi-Vallebona P: **Transcriptional activity of human endogenous retroviruses in human peripheral blood mononuclear cells**. *Biomed Res Int* 2015, **2015**.

18. Voisset C, Weiss RA, Griffiths DJ: **Human RNA "rumor" viruses: the search for novel human retroviruses in chronic disease.** *Microbiol Mol Biol Rev* 2008, **72**:157–96, table of contents.

19. Jern P, Coffin JM: **Effects of Retroviruses on Host Genome Function**. *Annu Rev Genet* 2008, **42**:709–732.

20. Christensen T: **Human endogenous retroviruses in neurologic disease**. *Apmis* 2016, **124**:116–126.

21. Grandi N, Cadeddu M, Blomberg J, Tramontano E: **Contribution of type W human endogenous retrovirus to the human genome: characterization of HERV-W proviral insertions and processed pseudogenes**. *Retrovirology* 2016, **13**:1–25.

22. Perron H, Garson JA, Bedin F, Beseme F, Paranhos-Baccala G, Komurian-Pradel F, Mallet F, Tuke PW, Voisset C, Blond JL, Lalande B, Seigneurin JM, Mandrand B: **Molecular identification of a novel retrovirus repeatedly isolated from patients with multiple sclerosis. The Collaborative Research Group on Multiple Sclerosis.** *Proc Natl Acad Sci U S A* 1997, **94**:7583–7588.

23. Blond JL, Besème F, Duret L, Bouton O, Bedin F, Perron H, Mandrand B, Mallet F: **Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family.** *J Virol* 1999, **73**:1175–85.

24. Blond JL, Lavillette D, Cheynet V, Bouton O, Oriol G, Chapel-Fernandes S, Mandrand B, Mallet F, Cosset FL: **An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor.** *J Virol* 2000, **74**:3321–9.

25. Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, Lavallie E, Tang X, Edouard P, Howes S, Jr JCK, Mccoy JM: **Syncytin is a captive retroviral envelope protein involved**. *Nature* 2000, **403**(February):785–789.

26. Cheng Y-H: **Isolation and Characterization of the Human Syncytin Gene Promoter**. *Biol Reprod* 2003, **70**:694–701.

27. Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, Lucotte G, Duret L, Mandrand B: **The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology.** *Proc Natl Acad Sci U S A* 2004, **101**:1731–6.

28. Bonnaud B, Bouton O, Oriol G, Cheynet V, Duret L, Mallet F: **Evidence of selection on the domesticated ERVWE1 env retroviral element involved in placentation.** *Mol Biol Evol* 2004, **21**:1895–901.

29. Prudhomme S, Oriol G, Mallet F: **A retroviral promoter and a cellular enhancer define a bipartite element which controls env ERVWE1 placental expression.** *J Virol* 2004, **78**:12157–12168.

30. Cheynet V, Ruggieri A, Oriol G, Blond J-L, Boson B, Vachot L, Verrier B, Cosset F-L, Mallet F: **Synthesis, assembly, and processing of the Env ERVWE1/syncytin human endogenous retroviral envelope.** *J Virol* 2005, **79**:5585–93.

31. Gimenez J, Mallet F: **ERVWE1 (endogenous retroviral family W, Env(C7), member 1)**. *Atlas Genet Cytogenet Oncol Haematol* 2008, **12**:134–148.

32. Mayer J, Meese E: **Human endogenous retroviruses in the primate lineage and their influence on host genomes.** *Cytogenet Genome Res* 2005, **110**:448–56.

33. Pavlícek A, Paces J, Elleder D: **Processed Pseudogenes of Human Endogenous Retroviruses Generated by LINEs: Their Integration, Stability, and Distribution**. *Genome Res* 2002, **12**:391–399.

34. Costas J: **Characterization of the intragenomic spread of the human endogenous retrovirus family HERV-W.** *Mol Biol Evol* 2002, **19**:526–33.

35. Lavie L, Maldener E, Brouha B, Meese EU, Mayer J: **The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity.** *Genome Res* 2004, **14**:2253–60.

36. Voisset C, Bouton O, Bedin F, Duret L, Mandrand B, Mallet F, Paranhos-Baccala G: **Chromosomal distribution and coding capacity of the human endogenous retrovirus HERV-W family.** *AIDS Res*

Hum Retroviruses 2000, **16**:731–40.

37. Perron C, Geny A, Laurent C, Mouriquand J, Pellat J, Perret J, Seigneurin J: **Leptomeningeal cell line from multiple sclerosis with reverse transcriptase activity and viral particles.** *Res Virol* 1989, **140**:551–61.

38. Perron H, Lalande B, Gratacap B, Laurent A, Genoulaz O, Geny C, Mallaret M, Schuller E, Stoebner P, Seigneurin J: **Isolation of retrovirus from patients with multiple sclerosis**. *Lancet* 1991, **337**:862–3.

39. Knerr I, Huppertz B, Weigel C, Dötsch J, Wich C, Schild RL, Beckmann MW, Rascher W: **Endogenous retroviral syncytin: Compilation of experimental research on syncytin and its possible role in normal and disturbed human placentogenesis**. *Mol Hum Reprod* 2004, **10**:581–588.

40. Frendo J-L, Olivier D, Cheynet V, Blond J-L, Bouton O, Vidaud M, Rabreau M, Evain-Brion D, Mallet F: **Direct involvement of HERV-W Env glycoprotein in human trophoblast cell fusion and differentiation.** *Mol Cell Biol* 2003, **23**:3566–74.

41. Lavillette D, Marin M, Ruggieri A, Mallet F, Cosset FL, Kabat D: **The envelope glycoprotein of human endogenous retrovirus type W uses a divergent family of amino acid transporters/cell surface receptors**. *J Virol* 2002, **76**:6442–6452.

42. Malassiné A, Handschuh K, Tsatsaris V, Gerbaud P, Cheynet V, Oriol G, Mallet F, Evain-Brion D: **Expression of HERV-W Env glycoprotein (syncytin) in the extravillous trophoblast of first trimester human placenta**. *Placenta* 2005, **26**:556–562.

43. Huang Q, Li J, Wang F, Oliver MT, Tipton T, Gao Y, Jiang S-W: **Syncytin-1 modulates placental trophoblast cell proliferation by promoting G1/S transition.** *Cell Signal* 2013, **25**:1027–35.

44. Keryer G, Alsat E, Tasken K, Evain-Brion D: **Cyclic AMP-dependent protein kinases and human trophoblast cell differentiation in vitro.** *J Cell Sci* 1998, **111 (Pt 7**:995–1004.

45. Huang Q, Chen H, Wang F, Brost BC, Li J, Li Z, Gao Y, Gao Y, Jiang SW: **Reduced syncytin-1 expression in choriocarcinoma BeWo cells activates the calpain1-AIF-mediated apoptosis, implication for preeclampsia**. *Cell Mol Life Sci* 2014, **71**:3151–3164.

46. Mangeney M, Heidmann T: **Tumor cells expressing a retroviral envelope escape immune rejection in vivo.** *Proc Natl Acad Sci U S A* 1998, **95**(December):14920–14925.

47. Blaise S, Mangeney M, Heidmann T: **The envelope of Mason-Pfizer monkey virus has immunosuppressive properties**. *J Gen Virol* 2001, **82**:1597–1600.

48. Mangeney M, Renard M, Schlecht-Louf G, Bouallaga I, Heidmann O, Letzelter C, Richaud A, Ducos B, Heidmann T: **Placental syncytins: Genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins.** *Proc Natl Acad Sci U S A* 2007, **104**:20534–20539.

49. Tolosa JM, Schjenken JE, Clifton VL, Vargas A, Barbeau B, Lowry P, Maiti K, Smith R: **The endogenous retroviral envelope protein syncytin-1 inhibits LPS/PHA-stimulated cytokine responses in human blood and is sorted into placental exosomes**. *Placenta* 2012, **33**:933–941.

50. Li F, Karlsson H: **Expression and regulation of human endogenous retrovirus W elements**. *Apmis* 2016, **124**:52–66.

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

52. Kim HS, Ahn K, Kim DS: **Quantitative expression of the HERV-W env gene in human tissues**. *Arch Virol* 2008, **153**:1587–1591.

53. Hu L, Hornung D, Kurek R, Ostman H, Helen O, Blomberg J, Bergqvist A: **Expression of human endogenous gammaretroviral sequences in endometriosis and ovarian cancer.** *AIDS Res Hum Retroviruses* 2006, **22**:551–7.

54. Hu L, Östman H, Bergqvist A, Blomberg J: **Physiological expression of human endogenous gammaretrovirus-like RNA in female reproductive tissues**. 2007.

55. Schön U, Seifarth W, Baust C, Hohenadl C, Erfle V, Leib-Mösch C: **Cell Type-Specific Expression and Promoter Activity of Human Endogenous Retroviral Long Terminal Repeats**. *Virology* 2001, **279**:280–291.

56. Nellåker C, Wållgren U, Karlsson H: **Molecular beacon-based temperature control and automated analyses for improved resolution of melting temperature analysis using SYBR I Green chemistry**. *Clin Chem* 2007, **53**:98–103.

57. Nellåker C, Li F, Uhrzander F, Tyrcha J, Karlsson H: **Expression profiling of repetitive elements by melting temperature analysis: variation in HERV-W gag expression across human individuals and tissues.** *BMC Genomics* 2009, **10**:532.

58. Roland CS, Hu J, Ren CE, Chen H, Li J, Varvoutis MS, Leaphart LW, Byck DB, Zhu X, Jiang SW: **Morphological changes of placental syncytium and their implications for the pathogenesis of preeclampsia**. *Cell Mol Life Sci* 2016, **73**:365–376.

59. Vargas A, Toufaily C, LeBellego F, Rassart É, Lafond J, Barbeau B: **Reduced expression of both syncytin 1 and syncytin 2 correlates with severity of preeclampsia.** *Reprod Sci* 2011, **18**:1085–91.

60. Lee X, Keith JC, Stumm N, Moutsatsos I, McCoy JM, Crum CP, Genest D, Chin D, Ehrenfels C, Pijnenborg R, Van Assche FA, Mi S: **Downregulation of placental syncytin expression and abnormal protein localization in pre-eclampsia**. *Placenta* 2001, **22**:808–812.

61. Knerr I, Beinder E, Rascher W: **Syncytin, a novel human endogenous retroviral gene in human placenta: Evidence for its dysregulation in preeclampsia and HELLP syndrome**. *Am J Obstet Gynecol* 2002, **186**:210–213.

62. Holder BS, Tower CL, Abrahams VM, Aplin JD: **Syncytin 1 in the human placenta**. *Placenta* 2012, **33**:460–466.

63. Zhuang X-W, Li J, Brost BC, Xia X-Y, Chen H Bin, Wang C-X, Jiang S-W: **Decreased expression and altered methylation of syncytin-1 gene in human placentas associated with preeclampsia.** *Curr Pharm Des* 2014, **20**:1796–802.

64. Knerr I, Weigel C, Linnemann K, Dotsch J, Meissner U, Fusch C, Rascher W: **Transcriptional effects of hypoxia on fusiogenic syncytin and its receptor ASCT2 in human cytotrophoblast BeWo cells and in ex vivo perfused placental cotyledons**. *Am J Obs Gynecol* 2003, **189**:583–588.

65. Muir a, Lever a, Moffett a: **Expression and functions of human endogenous retroviruses in the placenta: an update.** *Placenta* 2004, **25 Suppl A**:S16-25.

66. Frendo JL, Vidaud M, Guibourdenche J, Luton D, Muller F, Bellet D, Giovagrandi Y, Tarrade A, Porquet D, Blot P, Evain-Brion D: **Defect of villous cytotrophoblast differentiation into syncytiotrophoblast in Down's syndrome**. *J Clin Endocrinol Metab* 2000, **85**:3700–3707.

67. Massin N, Frendo JL, Guibourdenche J, Luton D, Giovangrandi Y, Muller F, Vidaud M, Evain-Brion D: **Defect of syncytiotrophoblast formation and human chorionic gonadotropin expression in Down's syndrome**. *Placenta* 2001, **22**(SUPPL.1).

68. Malassiné A, Frendo JL, Evain-Brion D: **Trisomy 21- Affected placentas highlight prerequisite factors for human trophoblast fusion and differentiation**. *Int J Dev Biol* 2010, **54**:475–482.

69. Frendo JL, Thérond P, Bird T, Massin N, Muller F, Guibourdenche J, Luton D, Vidaud M, Anderson WB, Evain-Brion D: **Overexpression of copper zinc superoxide dismutase impairs human trophoblast cell fusion and differentiation**. *Endocrinology* 2001, **142**:3638–3648.

70. Ruebner M, Strissel PL, Langbein M, Fahlbusch F, Wachter DL, Faschingbauer F, Beckmann MW, Strick R: **Impaired cell fusion and differentiation in placentae from patients with intrauterine growth restriction correlate with reduced levels of HERV envelope genes**. *J Mol Med* 2010, **88**:1143–

1156.

71. Zhou H, Li J, Podratz KC, Tipton T, Marzolf S, Chen H Bin, Jiang S-W: **Hypomethylation and activation of syncytin-1 gene in endometriotic tissue.** *Curr Pharm Des* 2014, **20**:1786–95.

72. Galli UM, Sauter M, Lecher B, Maurer S, Herbst H, Roemer K, Mueller-Lantzsch N: **Human endogenous retrovirus rec interferes with germ cell development in mice and may cause carcinoma in situ, the predecessor lesion of germ cell tumors.** *Oncogene* 2005, **24**:3223–8.

73. Flockerzi A, Ruggieri A, Frank O, Sauter M, Maldener E, Kopper B, Wullich B, Seifarth W, Müller-Lantzsch N, Leib-Mösch C, Meese E, Mayer J: **Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project.** *BMC Genomics* 2008, **9**:354.

74. Kassiotis G: **Endogenous retroviruses and the development of cancer.** *J Immunol* 2014, **192**:1343–9.

75. Menendez L, Benigno BB, McDonald JF: **L1 and HERV-W retrotransposons are hypomethylated in human ovarian carcinomas.** *Mol Cancer* 2004, **3**:12.

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

77. Díaz-Carballo D, Acikelli AH, Klein J, Jastrow H, Dammann P, Wyganowski T, Guemues C, Gustmann S, Bardenheuer W, Malak S, Tefett NS, Khosrawipour V, Giger-Pabst U, Tannapfel A, Strumberg D: **Therapeutic potential of antiviral drugs targeting chemorefractory colorectal adenocarcinoma cells overexpressing endogenous retroviral elements**. *J Exp Clin Cancer Res* 2015, **34**:81.

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

79. Schön U, Seifarth W, Baust C, Hohenadl C, Erfle V, Leib-Mösch C: **Cell type-specific expression and promoter activity of human endogenous retroviral long terminal repeats.** *Virology* 2001, **279**:280–91.

80. Hu L, Hedborg F, Blomberg J: **Selective activation of HERVW RNA expression during culture of a neuroblastoma cell line**. 2007.

81. Hu L, Uzhameckis D, Hedborg F, Blomberg J: **Dynamic and selective HERV RNA expression in neuroblastoma cells subjected to variation in oxygen tension and demethylation**. *Apmis* 2016, **124**:140–149.

82. Li S, Liu ZC, Yin SJ, Chen YT, Yu HL, Zeng J, Zhang Q, Zhu F: **Human endogenous retrovirus W family envelope gene activates the small conductance Ca2+-activated K+ channel in human neuroblastoma cells through CREB**. *Neuroscience* 2013, **247**:164–174.

83. Gimenez J, Montgiraud C, Pichon J-P, Bonnaud B, Arsac M, Ruel K, Bouton O, Mallet F: **Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control.** *Nucleic Acids Res* 2010, **38**:2229–46.

84. Maliniemi P, Vincendeau M, Mayer J, Frank O, Hahtola S, Karenko L, Carlsson E, Mallet F, Seifarth W, Leib-Mösch C, Ranki A: **Expression of human endogenous retrovirus-w including syncytin-1 in cutaneous T-cell lymphoma.** *PLoS One* 2013, **8**:e76281.

85. Yu H, Liu T, Zhao Z, Chen Y, Zeng J, Liu S, Zhu F: **Mutations in 3′-long terminal repeat of HERV-W family in chromosome 7 upregulate syncytin-1 expression in urothelial cell carcinoma of the bladder through interacting with c-Myb**. *Oncogene* 2014, **33**:3947–3958.

86. Stauffer Y, Theiler G, Sperisen P, Lebedev Y, Jongeneel C V: **Digital expression profiles of human**

endogenous retroviral families in normal and cancerous tissues. *Cancer Immun* 2004, **2**:1–18.

87. Ruprecht K, Mayer J, Sauter M, Roemer K, Mueller-Lantzsch N: **Endogenous retroviruses and cancer.** *Cell Mol Life Sci* 2008, **65**:3366–82.

88. Balada E, Ordi-Ros J, Vilardell-Tarrés M: **Molecular mechanisms mediated by Human Endogenous Retroviruses (HERVs) in autoimmunity**. *Rev Med Virol* 2009, **19**:273–286.

89. Balada E, Vilardell-Tarrés M, Ordi-Ros J: **Implication of human endogenous retroviruses in the development of autoimmune diseases.** *Int Rev Immunol* 2010, **29**:351–70.

90. Trela M, Nelson PN, Rylance PB: **The role of molecular mimicry and other factors in the association of Human Endogenous Retroviruses and autoimmunity.** *APMIS* 2016, **124**:88–104.

91. Brodziak A, Zi E, Nowakowska-zajdel E, Kokot T, Klakla K: **The role of human endogenous retroviruses in autoimmune diseases**. 2011, **18**.

92. Volkman HE, Stetson DB: **The enemy within: endogenous retroelements and autoimmune disease**. *Nat Immunol* 2014, **15**:415–422.

93. Hurst TP, Magiorkinis G: **Activation of the innate immune response by endogenous retroviruses.** *J Gen Virol* 2015, **96**(Pt 6):1207–1218.

94. Query CC, Keene JD: **A human autoimmune protein associated with U1 RNA contains a region of homology that is cross-reactive with retroviral p30gag antigen**. *Cell* 1987, **51**:211–220.

95. Talal N, Flescher E, Dang H: **Are endogenous retroviruses involved in human autoimmune disease?** *J Autoimmun* 1992, **5 Suppl A**:61–66.

96. Nelson PN, Lever a M, Bruckner FE, Isenberg D a, Kessaris N, Hay FC: **Polymerase chain reaction fails to incriminate exogenous retroviruses HTLV-I and HIV-1 in rheumatological diseases although a minority of sera cross react with retroviral antigens.** *Ann Rheum Dis* 1994, **53**:749–54.

97. Mason AL, Xu L, Guo L, Garry RF: **Retroviruses in autoimmune liver disease: genetic or environmental agents?** *Arch Immunol Ther Exp* 1999, **47**:289–297.

98. Chuong EB, Elde NC, Feschotte C: **Regulatory evolution of innate immunity through co-option of endogenous retroviruses**. *Science (80-)* 2016, **351**:1083–1087.

99. Sun B, Hu L, Luo ZY, Chen XP, Zhou HH, Zhang W: **DNA methylation perspectives in the pathogenesis of autoimmune diseases**. *Clin Immunol* 2016, **164**:21–27.

100. Antony JM, Deslauriers AM, Bhat RK, Ellestad KK, Power C: **Human endogenous retroviruses and multiple sclerosis: innocent bystanders or disease determinants?** *Biochim Biophys Acta* 2011, **1812**:162–76.

101. Lafon M, Jouvin-Marche E, Marche PN, Perron H: **Human viral superantigens: to be or not to be transactivated?** *Trends Immunol* 2002:238–239.

102. Christensen T: **Association of human endogenous retroviruses with multiple sclerosis and possible interactions with herpes viruses.** *Rev Med Virol* 2005, **15**:179–211.

103. Perron H, Bernard C, Bertrand J-B, Lang A, Popa I, Sanhadji K, Portoukalian J: **Endogenous retroviral genes, Herpesviruses and gender in Multiple Sclerosis.** *J Neurol Sci* 2009, **286**:65–72.

104. Krone B, Grange JM: **Multiple Sclerosis: Are Protective Immune Mechanisms Compromised by a Complex Infectious Background?** *Autoimmune Dis* 2011, **2011**:1–8.

105. Libbey JE, Cusick MF, Fujinami RS: **Role of Pathogens in Multiple Sclerosis.** *Int Rev Immunol* 2013, **33**(July 2013):1–18.

106. Morandi E, Tarlinton RE, Gran B: **Multiple sclerosis between genetics and infections: Human endogenous retroviruses in monocytes and macrophages**. *Front Immunol* 2015, **6**(DEC):1–6.

107. Alliel PM, Perin JP, Belliveau J, Pierig R, Nussbaum JL, Rieger F: **[Endogenous retroviral**

sequences analogous to that of the new retrovirus MSRV associated with multiple sclerosis (part 1)]. *CRAcad Sci III* 1998, **321**:495–499.

108. Alliel PM, Perin JP, Pierig R, Rieger F: **An endogenous retrovirus with nucleic acid sequences similar to those of the multiple sclerosis associated retrovirus at the human T-cell receptor alpha, delta gene locus.** *Cell Mol Biol (Noisy-le-grand)* 1998.

109. Haahr S, Sommerlund M, Møller-Larsen A, Nielsen R, Hansen H: **Just another dubious virus in cells from a patient with multiple sclerosis?** *Lancet* 1991:863–864.

110. Perron H, Firouzi R, Tuke P, Garson JA, Michel M, Beseme F, Bedin F, Mallet F, Marcel E, Seigneurin JM, Mandrand B: **Cell cultures and associated retroviruses in multiple sclerosis**. *Acta Neurol Scand Suppl* 1997, **169**:22–31.

111. Komurian-Pradel F, Paranhos-Baccala G, Bedin F, Ounanian-Paraz A, Sodoyer M, Ott C, Rajoharison A, Garcia E, Mallet F, Mandrand B, Perron H: **Molecular cloning and characterization of MSRV-related sequences associated with retrovirus-like particles.** *Virology* 1999, **260**:1–9.

112. Mameli G, Astone V, Arru G, Marconi S, Lovato L, Serra C, Sotgiu S, Bonetti B, Dolei A: **Brains and peripheral blood mononuclear cells of multiple sclerosis (MS) patients hyperexpress MS-associated retrovirus/HERV-W endogenous retrovirus, but not Human herpesvirus 6.** *J Gen Virol* 2007, **88**(Pt 1):264–274.

113. Mameli G, Poddighe L, Astone V, Delogu G, Arru G, Sotgiu S, Serra C, Dolei A: **Novel reliable real-time PCR for differential detection of MSRVenv and syncytin-1 in RNA and DNA from patients with multiple sclerosis**. *J Virol Methods* 2009, **161**:98–106.

114. Dolei A, Perron H: **The multiple sclerosis-associated retrovirus and its HERV-W endogenous family: a biological interface between virology, genetics, and immunology in human physiology and disease.** *J Neurovirol* 2009, **15**:4–13.

115. Garcia-Montojo M, Dominguez-Mozo M, Arias-Leal A, Garcia-Martinez Á, de las Heras V, Casanova I, Faucard R, Gehin N, Madeira A, Arroyo R, Curtin F, Alvarez-Lafuente R, Perron H: **The DNA Copy Number of Human Endogenous Retrovirus-W (MSRV-Type) Is Increased in Multiple Sclerosis Patients and Is Influenced by Gender and Disease Severity**. *PLoS One* 2013, **8**.

116. Blomberg J, Ushameckis D, Jern P: *Evolutionary Aspects of Human Endogenous Retroviral Sequences (HERVs) and Disease*. Landes Bioscience; 2000.

117. Voisset C, Weiss R a, Griffiths DJ: **Human RNA "rumor" viruses: the search for novel human retroviruses in chronic disease.** *Microbiol Mol Biol Rev* 2008, **72**:157–196, table of contents.

118. Ruprecht K, Gronen F, Sauter M, Best B, Rieckmann P, Mueller-Lantzsch N: **Lack of immune responses against multiple sclerosis-associated retrovirus/human endogenous retrovirus W in patients with multiple sclerosis.** *J Neurovirol* 2008, **14**:143–51.

119. Schmitt K, Richter C, Backes C, Meese E, Ruprecht K, Mayer J: **Comprehensive analysis of human endogenous retrovirus group HERV-W locus transcription in multiple sclerosis brain lesions by high-throughput amplicon sequencing.** *J Virol* 2013, **87**:13837–52.

120. Perron H, Germi R, Bernard C, Garcia-Montojo M, Deluen C, Farinelli L, Faucard R, Veas F, Stefas I, Fabriek BO, Van-Horssen J, Van-der-Valk P, Gerdil C, Mancuso R, Saresella M, Clerici M, Marcel S, Creange A, Cavaretta R, Caputo D, Arru G, Morand P, Lang AB, Sotgiu S, Ruprecht K, Rieckmann P, Villoslada P, Chofflon M, Boucraut J, Pelletier J, et al.: **Human endogenous retrovirus type W envelope expression in blood and brain cells provides new insights into multiple sclerosis disease.** *Mult Scler* 2012, **18**:1721–36.

121. Grandi N, Cadeddu M, Blomberg J, Tramontano E: **Contribution of type W human endogenous retrovirus to the human genome: characterization of HERV-W proviral insertions and processed pseudogenes**. *Retrovirology* 2016, **accepted**.

122. Zawada M, Liwén I, Pernak M, Januszkiewicz-Lewandowska D, Nowicka-Kujawska K,

Rembowska J, Lewandowski K, Hertmanowska H, Wender M, Nowak J: **MSRV pol sequence copy number as a potential marker of multiple sclerosis**. *Pol J Pharmacol* 2003, **55**:869–875.

123. Nowak J, Januszkiewicz D, Pernak M, Liwé I, Zawada M, Rembowska J, Nowicka K, Lewandowski K, Hertmanowska H, Wender M: **Multiple sclerosis–associated virus-related pol sequences found both in multiple sclerosis and healthy donors are more frequently expressed in multiple sclerosis patients**. *J Neurovirol* 2003, **9**:112–117.

124. Johnston JB, Silva C, Holden J, Warren KG, Clark AW, Power C: **Monocyte activation and differentiation augment human endogenous retrovirus expression: Implications for inflammatory brain diseases**. *Ann Neurol* 2001, **50**:434–442.

125. Dolei a., Serra C, Mameli G, Pugliatti M, Sechi G, Cirotto MC, Rosati G, Sotgiu S: **Multiple sclerosis-associated retrovirus (MSRV) in Sardinian MS patients.** *Neurology* 2002, **58**:471–3.

126. Nowak J, Januszkiewicz D, Pernak M, Liwen II, Zawada M, Rembowska J, Nowicka K, Lewandowski K, Hertmanowska H, Wender M: **Multiple sclerosis-associated virus-related pol sequences found both in multiple sclerosis and healthy donors are more frequently expressed in multiple sclerosis patients**. *J Neurovirol* 2003, **9**:112–117.

127. Garson J, Tuke P, Giraud P, Paranhos-Baccala G, Perron H: **Detection of virion-associated MSRV-RNA in serum of patients with multiple sclerosis**. *Lancet* 1998, **351**.

128. Antony JM, van Marle G, Opii W, Butterfield DA, Mallet F, Yong VW, Wallace JL, Deacon RM, Warren K, Power C: **Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination.** *Nat Neurosci* 2004, **7**:1088–95.

129. Antony JM, Zhu Y, Izad M, Warren KG, Vodjgani M, Mallet F, Power C: **Comparative Expression of Human Endogenous Retrovirus-W Genes in Multiple Sclerosis**. *AIDS Res Hum Retroviruses* 2007, **23**:1251–1256.

130. Arru G, Mameli G, Astone V, Serra C, Huang Y-M, Link H, Fainardi E, Castellazzi M, Granieri E, Fernandez M, Villoslada P, Fois ML, Sanna A, Rosati G, Dolei A, Sotgiu S: **Multiple Sclerosis and HERV-W/MSRV: A Multicentric Study.** *Int J Biomed Sci* 2007, **3**:292–7.

131. Mameli G, Astone V, Arru G, Marconi S, Lovato L, Serra C, Sotgiu S, Bonetti B, Dolei A: **Brains and peripheral blood mononuclear cells of multiple sclerosis (MS) patients hyperexpress MS-associated retrovirus/HERV-W endogenous retrovirus, but not Human herpesvirus 6.** *J Gen Virol* 2007, **88**(Pt 1):264–274.

132. Brudek T, Christensen T, Aagaard L, Petersen T, Hansen HJ, Møller-Larsen A: **B cells and monocytes from patients with active multiple sclerosis exhibit increased surface expression of both HERV-H Env and HERV-W Env, accompanied by increased seroreactivity.** *Retrovirology* 2009, **6**:104.

133. Perron H, Lazarini F, Ruprecht K, Péchoux-Longin C, Seilhean D, Sazdovitch V, Créange A, Battail-Poirot N, Sibaï G, Santoro L, Jolivet M, Darlix J-L, Rieckmann P, Arzberger T, Hauw J-J, Lassmann H: **Human endogenous retrovirus (HERV)-W ENV and GAG proteins: physiological expression in human brain and pathophysiological modulation in multiple sclerosis lesions.** *J Neurovirol* 2005, **11**:23–33.

134. Van Horssen J, Van Der Pol S, Nijland P, Amor S, Perron H: **Human endogenous retrovirus W in brain lesions: Rationale for targeted therapy in multiple sclerosis**. *Mult Scler Relat Disord* 2016, **8**:11–18.

135. Perron H, Jouvin-Marche E, Michel M, Ounanian-Paraz A, Camelo S, Dumon A, Jolivet-Reynaud C, Marcel F, Souillet Y, Borel E, Gebuhrer L, Santoro L, Marcel S, Seigneurin JM, Marche PN, Lafon M: **Multiple Sclerosis Retrovirus Particles and Recombinant Envelope Trigger an Abnormal Immune Response in Vitro, by Inducing Polyclonal Vβ16 T-Lymphocyte Activation**. *Virology* 2001, **287**:321–332.

136. Rolland A, Jouvin-Marche E, Saresella M, Ferrante P, Cavaretta R, Créange A, Marche P, Perron

H: **Correlation between disease severity and in vitro cytokine production mediated by MSRV (Multiple Sclerosis associated RetroViral element) envelope protein in patients with multiple sclerosis**. *J Neuroimmunol* 2005, **160**:195–203.

137. Rolland a., Jouvin-Marche E, Viret C, Faure M, Perron H, Marche PN: **The Envelope Protein of a Human Endogenous Retrovirus-W Family Activates Innate Immunity through CD14/TLR4 and Promotes Th1-Like Responses**. *J Immunol* 2006, **176**:7636–7644.

138. Saresella M, Rolland A, Marventano I, Cavarretta R, Caputo D, Marche P, Perron H, Clerici M: **Multiple sclerosis-associated retroviral agent (MSRV)-stimulated cytokine production in patients with relapsing-remitting multiple sclerosis.** *Mult Scler* 2009, **15**:443–7.

139. Mameli G, Astone V, Khalili K, Serra C, Sawaya BE, Dolei A: **Regulation of the syncytin-1 promoter in human astrocytes by multiple sclerosis-related cytokines**. *Virology* 2007, **362**:120–130.

140. Kremer D, Schichel T, Förster M, Tzekova N, Bernard C, Van Der Valk P, Van Horssen J, Hartung HP, Perron H, Küry P: **Human endogenous retrovirus type W envelope protein inhibits oligodendroglial precursor cell differentiation**. *Ann Neurol* 2013, **74**:721–732.

141. Kremer D, Förster M, Schichel T, Göttle P, Hartung H-P, Perron H, Küry P: **The neutralizing antibody GNbAC1 abrogates HERV-W envelope protein-mediated oligodendroglial maturation blockade.** *Mult Scler* 2014:1–4.

142. Madeira A, Burgelin I, Perron H, Curtin F, Lang AB, Faucard R: **MSRV envelope protein is a potent, endogenous and pathogenic agonist of human toll-like receptor 4: Relevance of GNbAC1 in multiple sclerosis treatment**. *J Neuroimmunol* 2016, **291**:29–38.

143. Petersen T, Møller-Larsen A, Thiel S, Brudek T, Hansen TK, Christensen T: **Effects of interferon-beta therapy on innate and adaptive immune responses to the human endogenous retroviruses HERV-H and HERV-W, cytokine production, and the lectin complement activation pathway in multiple sclerosis.** *J Neuroimmunol* 2009, **215**:108–16.

144. Mameli G, Serra C, Astone V, Castellazzi M, Poddighe L, Fainardi E, Neri W, Granieri E, Dolei A: **Inhibition of multiple-sclerosis-associated retrovirus as biomarker of interferon therapy.** *J Neurovirol* 2008, **14**:73–7.

145. Mameli G, Cossu D, Cocco E, Frau J, Marrosu MG, Niegowska M, Sechi LA: **Epitopes of HERV-Wenv induce antigen-specific humoral immunity in multiple sclerosis patients**. *J Neuroimmunol* 2015, **280**:66–68.

146. Firouzi R, Rolland A, Michel M, Jouvin-Marche E, Hauw J, Malcus-Vocanson C, Lazarini F, Gebuhrer L, Seigneurin J, Touraine J, Sanhadji K, Marche P, Perron H: **Multiple sclerosis–associated retrovirus particles cause T lymphocyte–dependent death with brain hemorrhage in humanized SCID mice model**. *J Neurovirol* 2003, **9**:79–93.

147. Antony JM, Ellestad KK, Hammond R, Imaizumi K, Mallet F, Warren KG, Power C: **The human endogenous retrovirus envelope glycoprotein, syncytin-1, regulates neuroinflammation and its receptor expression in multiple sclerosis: a role for endoplasmic reticulum chaperones in astrocytes**. *J Immunol* 2007, **179**:1210–1224.

148. Perron H, Dougier-Reynaud H-L, Lomparski C, Popa I, Firouzi R, Bertrand J-B, Marusic S, Portoukalian J, Jouvin-Marche E, Villiers CL, Touraine J-L, Marche PN: **Human endogenous retrovirus protein activates innate immunity and promotes experimental allergic encephalomyelitis in mice.** *PLoS One* 2013, **8**:e80128.

149. Nellaker C, Yao Y, Jones-Brando L, Mallet F, Yolken RH, Karlsson H: **Transactivation of elements in the human endogenous retrovirus W family by viral infection**. *Retrovirology* 2006, **3**:44.

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

151. Laufer G, Mayer J, Mueller BF, Mueller-Lantzsch N, Ruprecht K: **Analysis of transcribed human endogenous retrovirus W env loci clarifies the origin of multiple sclerosis-associated retrovirus env sequences.** *Retrovirology* 2009, **6**:37.

152. Roebke C, Wahl S, Laufer G, Stadelmann C, Sauter M, Mueller-Lantzsch N, Mayer J, Ruprecht K: **An N-terminally truncated envelope protein encoded by a human endogenous retrovirus W locus on chromosome Xq22.3.** *Retrovirology* 2010, **7**:69.

153. Mameli G, Poddighe L, Astone V, Delogu G, Arru G, Sotgiu S, Serra C, Dolei A: **Novel reliable real-time PCR for differential detection of MSRVenv and syncytin-1 in RNA and DNA from patients with multiple sclerosis**. *J Virol Methods* 2009, **161**:98–106.

154. Do Olival GS, Faria TS, Nali LHS, de Oliveira ACP, Casseb J, Vidal JE, Cavenaghi VB, Tilbery CP, Moraes L, Fink MCS, Sumita LM, Perron H, Romano CM: **Genomic analysis of ERVWE2 locus in patients with multiple sclerosis: Absence of genetic association but potential role of human endogenous retrovirus type W elements in molecular mimicry with myelin antigen**. *Front Microbiol* 2013, **4**(JUN):1–7.

155. García-Montojo M, de la Hera B, Varadé J, de la Encarnación A, Camacho I, Domínguez-Mozo M, Arias-Leal A, García-Martínez A, Casanova I, Izquierdo G, Lucas M, Fedetz M, Alcina A, Arroyo R, Matesanz F, Urcelay E, Alvarez-Lafuente R: **HERV-W polymorphism in chromosome X is associated with multiple sclerosis risk and with differential expression of MSRV.** *Retrovirology* 2014, **11**:2.

156. Varadé J, García-Montojo M, de la Hera B, Camacho I, García-Martnez MÁ, Arroyo R, Álvarez-Lafuente R, Urcelay E: **Multiple sclerosis retrovirus-like envelope gene: Role of the chromosome 20 insertion**. *BBA Clin* 2015, **3**:162–167.

157. Bhat RK, Ellestad KK, Wheatley BM, Warren R, Holt R a, Power C: **Age- and disease-dependent HERV-W envelope allelic variation in brain: association with neuroimmune gene expression.** *PLoS One* 2011, **6**:e19176.

158. Hon GM, Erasmus RT, Matsha T: **Multiple sclerosis-associated retrovirus and related human endogenous retrovirus-W in patients with multiple sclerosis: a literature review.** *J Neuroimmunol* 2013, **263**:8–12.

159. Garson JA, Huggett JF, Bustin SA, Pfaffl MW, Benes V, Vandesompele J, Shipley GL: **Unreliable Real-Time PCR Analysis of Human Endogenous Retrovirus-W (HERV-W) RNA Expression and DNA Copy Number in Multiple Sclerosis**. *AIDS Res Hum Retroviruses* 2009, **25**:377–378.

160. Bustin SA: **The reproducibility of biomedical research: Sleepers awake!** *Biomol Detect Quantif* 2014, **2**:35–42.

161. P. Ryan F: **Human Endogenous Retroviruses in Multiple Sclerosis: Potential for Novel Neuro-Pharmacological Research**. *Curr Neuropharmacol* 2011, **9**:360–369.

162. Curtin F, Perron H, Kromminga A, Porchet H, Lang AB: **Preclinical and early clinical development of GNbAC1, a humanized IgG4 monoclonal antibody targeting endogenous retroviral MSRV-Env protein**. *MAbs* 2015, **7**:265–275.

163. Curtin F, Perron H, Faucard R, Porchet H, Lang AB: **Treatment Against Human Endogenous Retrovirus: A Possible Personalized Medicine Approach for Multiple Sclerosis**. *Mol Diagn Ther* 2015, **19**:255–265.

164. Gaudin P, Ijaz S, Tuke PW, Marcel F, Paraz a, Seigneurin JM, Mandrand B, Perron H, Garson J a: **Infrequency of detection of particle-associated MSRV/HERV-W RNA in the synovial fluid of patients with rheumatoid arthritis.** *Rheumatology (Oxford)* 2000, **39**:950–4.

165. Bendiksen S, Martinez-Zubiavrra I, Tümmler C, Knutsen G, Elvenes J, Olsen E, Olsen R, Moens U: **Human Endogenous Retrovirus W Activity in Cartilage of Osteoarthritis Patients**. *Biomed Res Int* 2014:1–14.

166. Faucard R, Madeira A, Gehin N, Authier F-J, Panaite P-A, Lesage C, Burgelin I, Bertel M, Bernard

C, Curtin F, Lang AB, Steck AJ, Perron H, Kuntzer T, Créange A: **Human Endogenous Retrovirus and Neuroinflammation in Chronic Inflammatory Demyelinating Polyradiculoneuropathy**. *EBioMedicine* 2016, **6**:190–198.

167. Molès JP, Tesniere A, Guilhou JJ: **A new endogenous retroviral sequence is expressed in skin of patients with psoriasis**. *Br J Dermatol* 2005, **153**:83–89.

168. Hohenadl C, Germaier H, Walchner M, Hagenhofer M, Herrmann M, Stürzl M, Kind P, Hehlmann R, Erfle V, Leib-Mösch C: **Transcriptional activation of endogenous retroviral sequences in human epidermal keratinocytes by UVB irradiation.** *J Invest Dermatol* 1999, **113**:587–594.

169. Schanab O, Humer J, Gleiss A, Mikula M, Sturlan S, Grunt S, Okamoto I, Muster T, Pehamberger H, Waltenberger A: **Expression of human endogenous retrovirus K is stimulated by ultraviolet radiation in melanoma**. *Pigment Cell Melanoma Res* 2011, **24**:656–665.

170. de Sousa Nogueira MA, Biancardi Gavioli CF, Pereira NZ, de Carvalho GC, Domingues R, Aoki V, Sato MN: **Human endogenous retrovirus expression is inversely related with the up-regulation of interferon-inducible genes in the skin of patients with lichen planus**. *Arch Dermatological Res Sousa Nogueira MA, Biancardi Gavioli CF, Pereira NZ, al Hum Endog Retrovir Expr is inversely Relat with up-regulation Interf genes Ski patients with lichen planu* 2015, **307**:259–264.

171. Oluwole SO a, Yao Y, Conradi S, Kristensson K, Karlsson H: **Elevated levels of transcripts encoding a human retroviral envelope protein (syncytin) in muscles from patients with motor neuron disease.** *Amyotroph Lateral Scler* 2007, **8**:67–72.

172. Jeong B-H, Lee Y-J, Carp RI, Kim Y-S: **The prevalence of human endogenous retroviruses in cerebrospinal fluids from patients with sporadic Creutzfeldt-Jakob disease.** *J Clin Virol* 2010, **47**:136–142.

173. Balestrieri E, Arpino C, Matteucci C, Sorrentino R, Pica F, Alessandrelli R, Coniglio A, Curatolo P, Rezza G, Macciardi F, Garaci E, Gaudi S, Sinibaldi-Vallebona P: **HERVs Expression in Autism Spectrum Disorders**. *PLoS One* 2012, **7**.

174. Balestrieri E, Pitzianti M, Matteucci C, D'Agati E, Sorrentino R, Baratta A, Caterina R, Zenobi R, Curatolo P, Garaci E, Sinibaldi-Vallebona P, Pasini A: **Human endogenous retroviruses and ADHD.** *World J Biol Psychiatry* 2013(October):1–6.

175. Deb-Rinker P, Klempan TA, O'Reilly RL, Torrey EF, Singh SM: **Molecular characterization of a MSRV-like sequence identified by RDA from monozygotic twin pairs discordant for schizophrenia.** *Genomics* 1999, **61**:133–144.

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

177. Karlsson H, Schröder J, Bachmann S, Bottmer C, Yolken RH: **HERV-W-related RNA detected in plasma from individuals with recent-onset schizophrenia or schizoaffective disorder**. *Mol Psychiatry* 2003, **9**:12–13.

178. Yao Y, Schröder J, Nellåker C, Bottmer C, Bachmann S, Yolken RH, Karlsson H: **Elevated levels of human endogenous retrovirus-W transcripts in blood cells from patients with first episode schizophrenia**. *Genes, Brain Behav* 2008, **7**:103–112.

179. Huang W, Li S, Hu Y, Yu H, Luo F, Zhang Q, Zhu F: **Implication of the env gene of the human endogenous retrovirus W family in the expression of BDNF and DRD3 and development of recent-onset schizophrenia**. *Schizophr Bull* 2011, **37**:988–1000.

180. Qin C, Li S, Yan Q, Wang X, Chen Y, Zhou P, Lu M, Zhu F: **Elevation of Ser9 phosphorylation of GSK3β is required for HERV-W env-mediated BDNF signaling in human U251 cells**. *Neurosci Lett* 2016, **627**:84–91.

181. Perron H, Mekaoui L, Bernard C, Veas F, Stefas I, Leboyer M: **Endogenous Retrovirus Type W**

GAG and Envelope Protein Antigenemia in Serum of Schizophrenic Patients. *Biol Psychiatry* 2008, **64**:1019–1023.

182. Hegyi H: **GABBR1 has a HERV-W LTR in its regulatory region--a possible implication for schizophrenia.** *Biol Direct* 2013, **8**:5.

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

184. Weis S, Llenos IC, Sabunciyan S, Dulay JR, Isler L, Yolken R, Perron H: **Reduced expression of human endogenous retrovirus (HERV)-W GAG protein in the cingulate gyrus and hippocampus in schizophrenia, bipolar disorder, and depression.** *J Neural Transm* 2007, **114**:645–55.

185. Perron H, Hamdani N, Faucard R, Lajnef M, Jamain S, Daban-Huard C, Sarrazin S, LeGuen E, Houenou J, Delavest M, Moins-Teisserenc H, Moins-Teiserenc H, Bengoufa D, Yolken R, Madeira A, Garcia-Montojo M, Gehin N, Burgelin I, Ollagnier G, Bernard C, Dumaine A, Henrion A, Gombert A, Le Dudal K, Charron D, Krishnamoorthy R, Tamouza R, Leboyer M: **Molecular characteristics of Human Endogenous Retrovirus type-W in schizophrenia and bipolar disorder.** *Transl Psychiatry* 2012, **2**:e201.

186. Diem O, Schäffner M, Seifarth W, Leib-Mösch C: **Influence of antipsychotic drugs on human endogenous retrovirus (HERV) transcription in brain cells**. *PLoS One* 2012, **7**.

187. Gürtler C, Bowie AG: **Innate immune detection of microbial nucleic acids**. *Trends Microbiol* 2013:413–420.

188. Melder DC, Pankratz VS, Federspiel MJ: **Evolutionary Pressure of a Receptor Competitor Selects Different Subgroup A Avian Leukosis Virus Escape Variants with Altered Receptor Interactions**. *J Virol* 2003, **77**:10504–10514.

189. Spencer TE, Mura M, Gray CA, Griebel PJ, Palmarini M: **Receptor usage and fetal expression of ovine endogenous betaretroviruses: implications for coevolution of endogenous and exogenous retroviruses**. *J Virol* 2003, **77**:749–753.

190. An DS, Xie Ym, Chen IS: **Envelope gene of the human endogenous retrovirus HERV-W encodes a functional retrovirus envelope.** *J Virol* 2001, **75**:3488–3489.

191. Vincendeau M, Göttesdorfer I, Schreml JMH, Wetie AGN, Mayer J, Greenwood AD, Helfer M, Kramer S, Seifarth W, Hadian K, Brack-Werner R, Leib-Mösch C: **Modulation of human endogenous retrovirus (HERV) transcription during persistent and de novo HIV-1 infection**. *Retrovirology* 2015, **12**:27.

192. Hurst T, Pace M, Katzourakis A, Phillips R, Klenerman P, Frater J, Magiorkinis G: **Human endogenous retrovirus (HERV) expression is not induced by treatment with the histone deacetylase (HDAC) inhibitors in cellular models of HIV-1 latency.** *Retrovirology* 2016, **13**:10.

193. Uleri E, Mei A, Mameli G, Poddighe L, Serra C, Dolei A: **HIV Tat acts on endogenous retroviruses of the W family and this occurs via Toll-like receptor4: inference for neuroAIDS.** *AIDS* 2014(August):1–12.

194. Toufaily C, Landry S, Leib-Mosch C, Rassart E, Barbeau B: **Activation of LTRs from different human endogenous retrovirus (HERV) families by the HTLV-1 tax protein and T-cell activators**. *Viruses* 2011, **3**:2146–2159.

195. Garrison KE, Jones RB, Meiklejohn DA, Anwar N, Ndhlovu LC, Chapman JM, Erickson AL, Agrawal A, Spotts G, Hecht FM, Rakoff-Nahoum S, Lenz J, Ostrowski MA, Nixon DF: **T cell responses to human endogenous retroviruses in HIV-1 infection**. *PLoS Pathog* 2007, **3**:1617–1627.

196. Perron H, Suh M, Lalande B, Gratacap B, Laurent A, Stoebner P, Seigneurin JR: **Herpes simplex virus ICP0 and ICP4 immediate early proteins strongly enhance expression of a retrovirus harboured by a leptomeningeal cell line from a patient with multiple sclerosis**. *J Gen Virol* 1993,

74:65–72.

197. Ruprecht K, Obojes K, Wengel V, Gronen F, Kim KS, Perron H, Schneider-Schaulies J, Rieckmann P: **Regulation of human endogenous retrovirus W protein expression by herpes simplex virus type 1: implications for multiple sclerosis.** *J Neurovirol* 2006, **12**:65–71.

198. Lee WJ, Kwun HJ, Kim HS, Jang KL: **Activation of the human endogenous retrovirus W long terminal repeat by herpes simplex virus type 1 immediate early protein 1.** *Mol Cells* 2003, **15**:75–80.

199. Krone B, Oeffner F, Grange JM: **Is the risk of multiple sclerosis related to the "biography" of the immune system?** *J Neurol* 2009, **256**:1052–60.

200. Mameli G, Poddighe L, Mei A, Uleri E, Sotgiu S, Serra C, Manetti R, Dolei A: **Expression and activation by Epstein Barr virus of human endogenous retroviruses-W in blood cells and astrocytes: inference for multiple sclerosis.** *PLoS One* 2012, **7**:e44991.

201. Assinger A, Yaiw K-C, Göttesdorfer I, Leib-Mösch C, Söderberg-Nauclér C: **Human cytomegalovirus (HCMV) induces human endogenous retrovirus (HERV) transcription.** *Retrovirology* 2013, **10**:132.

202. Bergallo M, Galliano I, Montanari P, Gambarino S, Mareschi K, Ferro F, Fagioli F, Tovo PA, Ravanini P: **CMV induces HERV-K and HERV-W expression in kidney transplant recipients**. *J Clin Virol* 2015, **68**:28–31.

203. Yu C, Shen K, Lin M, Chen P, Lin C, Chang GD, Chen H: **GCMa regulates the syncytin-mediated trophoblastic fusion**. *J Biol Chem* 2002, **277**:50062–50068.

204. Li F, Nellåker C, Sabunciyan S, Yolken RH, Jones-Brando L, Johansson A-S, Owe-Larsson B, Karlsson H: **Transcriptional derepression of the ERVWE1 locus following influenza A virus infection.** *J Virol* 2014, **88**:4328–37.

205. Ponferrada VG, Mauck BS, Wooley DP: **The envelope glycoprotein of human endogenous retrovirus HERV-W induces cellular resistance to spleen necrosis virus**. *Arch Virol* 2003, **148**:659–675.

206. Machnik G, Klimacka-Nawrot E, Sypniewski D, Matczyńska D, Gałka S, Bednarek I, Okopień B: **Porcine endogenous retrovirus (PERV) infection of HEK-293 cell line alters expression of human endogenous retrovirus (HERV-W) sequences**. *Folia Biol (Czech Republic)* 2014, **60**:35–46.

207. Sassaman DM, Dombroski BA, Moran J V, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH: **Many human L1 elements are capable of retrotransposition.** *Nat Genet* 1997, **16**:37–43.

208. Richardson SR, Narvaiza I, Planegger RA, Weitzman MD, Moran J V.: **APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition**. *Elife* 2014, **3**:e02008.

209. Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, Heidmann T: **Paleovirology of "syncytins", retroviral env genes exapted for a role in placentation.** *Philos Trans R Soc Lond B Biol Sci* 2013, **368**:20120507.

210. Beck CR, Garcia-Perez JL, Badge RM, Moran J V: **LINE-1 Elements in Structural Variation and Disease**. *Annu Rev Genomics Hum Genet* 2011, **12**:187–215.

211. Hancks DC, Kazazian HH: **Active human retrotransposons: variation and disease.** *Curr Opin Genet Dev* 2012, **22**:191–203.

212. Sperber G, Airola T, Jern P, Blomberg J: **Automated recognition of retroviral sequences in genomic data--RetroTector.** *Nucleic Acids Res* 2007, **35**:4964–76.

213. Kent WJ: **BLAT - The BLAST-like alignment tool**. *Genome Res* 2002, **12**:656–664.

214. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ: **The**

UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* 2014, **42**(Database issue):D764-70.

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

216. Kent W, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome Res* 2002, **12**:996–1006.

217. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A: **Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data**. *Bioinformatics* 2012, **28**:1647–1649.

218. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol* 2013, **30**:772–80.

219. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: Molecular evolutionary genetics analysis version 6.0**. *Mol Biol Evol* 2013, **30**:2725–2729.

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

221. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira M a M, Kessing B, Pontius J, Roelke M, Rumpler Y, Schneider MPC, Silva A, O'Brien SJ, Pecon-Slattery J: **A molecular phylogeny of living primates**. *PLoS Genet* 2011, **7**:1–17.

222. Steiper ME, Young NM: **Primate molecular divergence dates.** *Mol Phylogenet Evol* 2006, **41**:384–94.

223. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: A sequence logo generator**. *Genome Res* 2004, **14**:1188–1190.

224. Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J: **tRNAdb 2009: compilation of tRNA sequences and tRNA genes.** *Nucleic Acids Res* 2009, **37**(Database issue):D159-62.

225. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33**(Database issue):D501-4.

226. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: **The UCSC Known Genes.** *Bioinformatics* 2006, **22**:1036–46.

227. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, et al.: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res* 2012, **22**:1760–74.

228. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, et al.: **Architecture of the human regulatory network derived from ENCODE data.** *Nature* 2012, **489**:91–100.

229. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, Rando OJ, Birney E, Myers RM, Noble WS, Snyder M, Weng Z: **Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.** *Genome Res* 2012, **22**:1798–812.

230. Sperber G, Lövgren A, Eriksson N-E, Benachenhou F, Blomberg J: **RetroTector online, a rational tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences.**

BMC Bioinformatics 2009, **10 Suppl 6**:S4.

231. Maydt J, Lengauer T: **Recco: recombination analysis using cost optimization.** *Bioinformatics* 2006, **22**:1064–71.

232. Villesen P, Aagaard L, Wiuf C, Pedersen FS: **Identification of endogenous retroviral reading frames in the human genome.** *Retrovirology* 2004, **1**:32.

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

234. Varmus HE: **Form and function of retroviral proviruses.** *Science* 1982, **216**:812–820.

235. Kim H-S: **Genomic impact, chromosomal distribution and transcriptional regulation of HERV elements.** *Mol Cells* 2012, **33**:539–44.

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

237. Khodosevich K, Lebedev Y, Sverdlov E: **Endogenous retroviruses and human evolution**. *Comp Funct Genomics* 2002:494–498.

238. Jern P, Coffin JM: **Effects of retroviruses on host genome function.** *Annu Rev Genet* 2008, **42**:709–32.

239. Schön U, Diem O, Leitner L, Günzburg WH, Mager DL, Salmons B, Leib-Mösch C: **Human endogenous retroviral long terminal repeat sequences as cell type-specific promoters in retroviral vectors.** *J Virol* 2009, **83**:12643–50.

240. Kowalski PE, Freeman JD, Mager DL: **Intergenic splicing between a HERV-H endogenous retrovirus and two adjacent human genes.** *Genomics* 1999, **57**:371–9.

241. Medstrand P, Landry JR, Mager DL: **Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans.** *J Biol Chem* 2001, **276**:1896–903.

242. Dunn CA, Medstrand P, Mager DL: **An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon.** *Proc Natl Acad Sci U S A* 2003, **100**:12841–6.

243. Jordan IK, Rogozin IB, Glazko G V, Koonin E V: **Origin of a substantial fraction of human regulatory sequences from transposable elements.** *Trends Genet* 2003, **19**:68–72.

244. van de Lagemaat LN, Landry J-R, Mager DL, Medstrand P: **Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions.** *Trends Genet* 2003, **19**:530–6.

245. Dunn CA, van de Lagemaat LN, Baillie GJ, Mager DL: **Endogenous retrovirus long terminal repeats as ready-to-use mobile promoters: the case of primate beta3GAL-T5.** *Gene* 2005, **364**:2–12.

246. Medstrand P, van de Lagemaat LN, Dunn CA, Landry J-R, Svenback D, Mager DL: **Impact of transposable elements on the evolution of mammalian gene regulation.** *Cytogenet Genome Res* 2005, **110**:342–52.

247. Sin HS, Huh JW, Kim DS, Kang DW, Min DS, Kim TH, Ha HS, Kim HH, Lee SY, Kim HS: **Transcriptional control of the HERV-H LTR element of the GSDML gene in human tissues and cancer cells**. *Arch Virol* 2006, **151**:1985–1994.

248. Piriyapongsa J, Polavarapu N, Borodovsky M, McDonald J: **Exonization of the LTR transposable elements in human genome.** *BMC Genomics* 2007, **8**:291.

249. Conley AB, Piriyapongsa J, Jordan IK: **Retroviral promoters in the human genome.** *Bioinformatics* 2008, **24**:1563–7.

250. Isbel L, Whitelaw E: **Endogenous retroviruses in mammals: an emerging picture of how ERVs**

modify expression of adjacent genes. *Bioessays* 2012, **34**:734–8.

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

252. Subramanian RP, Wildschutte JH, Russo C, Coffin JM: **Identification, characterization, and co1. Subramanian RP, Wildschutte JH, Russo C, Coffin JM. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. SM1. Retrovirology. 2011;8(**. *Retrovirology* 2011, **8**:90.

253. Pavlı́ček A, Pačes J, Zı́ka R, Hejnar J: **Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection**. *Gene* 2002, **300**:189–194.

254. Dangel AW, Mendoza AR, Menachery CD, Baker BJ, Daniel CM, Carroll MC, Wu LC, Yu CY: **The dichotomous size variation of human complement C4 genes is mediated by a novel family of endogenous retroviruses, which also establishes species-specific genomic patterns among Old World primates**. *Immunogenetics* 1994, **40**:425–436.

255. Kim HS, Takenaka O, Crow TJ: **Isolation and phylogeny of endogenous retrovirus sequences belonging to the HERV-W family in primates**. *J Gen Virol* 1999, **80**:2613–2619.

256. Voisset C, Blancher a, Perron H, Mandrand B, Mallet F, Paranhos-Baccalà G: **Phylogeny of a novel family of human endogenous retrovirus sequences, HERV-W, in humans and other primates.** *AIDS Res Hum Retroviruses* 1999, **15**:1529–1533.

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

258. Bowzard JB, Bennett RP, Krishna NK, Ernst SM, Rein A, Wills JW: **Importance of basic residues in the nucleocapsid sequence for retrovirus Gag assembly and complementation rescue.** *J Virol* 1998, **72**:9034–9044.

259. Malik HS, Eickbush TH: **Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons.** *J Virol* 1999, **73**:5186–5190.

260. Singleton TL, Levin HL: **A long terminal repeat retrotransposon of fission yeast has strong preferences for specific sites of insertion**. *Eukaryot Cell* 2002, **1**:44–55.

261. Jern P, Sperber GO, Ahlsén G, Blomberg J: **Sequence variability, gene structure, and expression of full-length human endogenous retrovirus H.** *J Virol* 2005, **79**:6325–6337.

262. Zsíros J, Jebbink MF, Lukashov V V, Voûte PA, Berkhout B: **Biased nucleotide composition of the genome of HERV-K related endogenous retroviruses and its evolutionary implications.** *J Mol Evol* 1999, **48**:102–11.

263. Mangeat B, Turelli P, Caron G, Friedli M, Perrin L, Trono D: **Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts.** *Nature* 2003, **424**:99–103.

264. Chiu Y-L, Greene WC: **The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements.** *Annu Rev Immunol* 2008, **26**:317–53.

265. Chiu Y-L, Witkowska HE, Hall SC, Santiago M, Soros VB, Esnault C, Heidmann T, Greene WC: **High-molecular-mass APOBEC3G complexes restrict Alu retrotransposition.** *Proc Natl Acad Sci U S A* 2006, **103**:15588–15593.

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

267. Medstrand P, Van De Lagemaat LN, Mager DL: **Retroelement distributions in the human genome: Variations associated with age and proximity to genes**. *Genome Res* 2002, **12**:1483–1495.

268. Li F, Nellåker C, Yolken RH, Karlsson H: **A systematic evaluation of expression of HERV-W**

elements; influence of genomic context, viral structure and orientation. *BMC Genomics* 2011, **12**:22.

269. Hadjiargyrou M, Delihas N: **The intertwining of transposable elements and non-coding RNAs.** *Int J Mol Sci* 2013, **14**:13307–28.

270. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860–921.

271. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, et al.: **The sequence of the human genome.** *Science* 2001, **291**:1304–1351.

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

273. Schiavetti F, Thonnard J, Colau D, Boon T, Coulie PG: **A human endogenous retroviral sequence encoding an antigen recognized on melanoma by cytolytic T lymphocytes**. *Cancer Res* 2002, **62**:5510–5516.

274. Blaise S, de Parseval N, Heidmann T: **Functional characterization of two newly identified Human Endogenous Retrovirus coding envelope genes.** *Retrovirology* 2005, **2**:19.

275. Christensen T: **HERVs in neuropathogenesis.** *J Neuroimmune Pharmacol* 2010, **5**:326–35.

276. Chance MR, Sagi I, Wirt MD, Frisbie SM, Scheuring E, Chen E, Bess Jr. JW, Henderson LE, Arthur LO, South TL, et al.: **Extended x-ray absorption fine structure studies of a retrovirus: equine infectious anemia virus cysteine arrays are coordinated to zinc**. *Proc Natl Acad Sci U S A* 1992, **89**:10041–10045.

277. Magiorkinis G, Blanco-Melo D, Belshaw R: **The decline of human endogenous retroviruses: extinction and survival**. *Retrovirology* 2015, **12**:1–12.

278. Baillie GJ, van de Lagemaat LN, Baust C, Mager DL: **Multiple groups of endogenous betaretroviruses in mice, rats, and other mammals.** *J Virol* 2004, **78**:5784–98.

279. Tarlinton RE, Meers J, Young PR: **Retroviral invasion of the koala genome.** *Nature* 2006, **442**:79–81.

280. Arnaud F, Caporale M, Varela M, Biek R, Chessa B, Alberti A, Golder M, Mura M, Zhang Y-P, Yu L, Pereira F, Demartini JC, Leymaster K, Spencer TE, Palmarini M: **A paradigm for virus-host coevolution: sequential counter-adaptations between endogenous and exogenous retroviruses.** *PLoS Pathog* 2007, **3**:e170.

281. Bannert N, Kurth R: **The evolutionary dynamics of human endogenous retroviral families.** *Annu Rev Genomics Hum Genet* 2006, **7**:149–73.

282. Feschotte C, Gilbert C: **Endogenous viruses: insights into viral evolution and impact on host biology**. *Nat Rev Genet* 2012, **13**:283–296.

283. Suntsova M, Garazha A, Ivanova A, Kaminsky D, Zhavoronkov A, Buzdin A: **Molecular functions of human endogenous retroviruses in health and disease**. *Cell Mol Life Sci* 2015, **72**:3653–3675.

284. Landry J-R, Rouhi A, Medstrand P, Mager DL: **The Opitz syndrome gene Mid1 is transcribed from a human endogenous retroviral promoter.** *Mol Biol Evol* 2002, **19**:1934–1942.

285. Conley AB, Piriyapongsa J, Jordan IK: **Retroviral promoters in the human genome**. *Bioinformatics*

2008, **24**:1563–1567.

286. Cohen CJ, Lock WM, Mager DL: **Endogenous retroviral LTRs as promoters for human genes: a critical assessment.** *Gene* 2009, **448**:105–14.

287. Yu HL, Zhao ZK, Zhu F: **The role of human endogenous retroviral long terminal repeat sequences in human cancer (Review)**. *Int J Mol Med* 2013, **32**:755–762.

288. Zhao M, Wang Z, Yung S, Lu Q: **Epigenetic dynamics in immunity and autoimmunity**. *Int J Biochem Cell Biol* 2015, **67**:65–74.

289. Esnault C, Maestre J, Heidmann T: **Human LINE retrotransposons generate processed pseudogenes.** *Nat Genet* 2000, **24**(april):363–367.

290. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: **Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes.** *Proc Natl Acad Sci U S A* 2003, **100**:11484–11489.

291. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ.** *Genome Res* 2003, **13**:103–107.

292. Blikstad V, Benachenhou F, Sperber GO, Blomberg J: **Evolution of human endogenous retroviral sequences: a conceptual account.** *Cell Mol Life Sci* 2008, **65**:3348–65.

293. Lavie L, Medstrand P, Schempp W, Mayer J, Meese E: **Human Endogenous Retrovirus Family Reconstruction of an Ancient Betaretrovirus in the Human Genome Human Endogenous Retrovirus Family HERV-K (HML-5): Status , Evolution , and Reconstruction of an Ancient Betaretrovirus in the Human Genome †**. 2004, **78**:8788–8798.

294. Flockerzi A, Burkhardt S, Schempp W, Meese E, Mayer J: **Human endogenous retrovirus HERV-K14 families: status, variants, evolution, and mobilization of other cellular sequences.** *J Virol* 2005, **79**:2941–2949.

295. Jurka J: **Repbase Update: A database and an electronic journal of repetitive elements**. *Trends Genet* 2000, **16**:418–420.

296. Cohen M, Larsson E: **Human endogenous retroviruses**. *BioEssays* 1988, **9**:191–196.

297. Kohany O, Gentles AJ, Hankus L, Jurka J: **Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor.** *BMC Bioinformatics* 2006, **7**:474.

298. Malik HS, Eickbush TH: **Phylogenetic analysis of Ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses**. *Genome Res* 2001, **11**:1187–1197.

299. Morandi E, Tanasescu R, Tarlinton RE, Constantinescu CS, Zhang W, Tench C, Gran B: **The association between human endogenous retroviruses and multiple sclerosis: A systematic review and meta-analysis.** *PLoS One* 2017, **12**:e0172415.