Università degli Studi di Cagliari

PhD DEGREE

Life, Environmental and Drug Sciences

Cycle XXXII

TITLE OF THE PHD THESIS

Characterization of the Human Endogenous Retrovirus (HERV) HML-6 group and identification of HERV differential expression in immunity

Scientific Disciplinary Sector(s)

Microbiologia BIO/19

PhD Student:	Maria Paola Pisano
Coordinator of the PhD Program	Prof. Simona Distinto
Supervisor	Prof. Enzo Tramontano

Final exam. Academic Year 2018 – 2019
Thesis defence: June - July 2020 Session

Summary

Abstract

Human Endogenous retroviruses (HERVs) are remnants of ancient retroviral infections that represent a large fraction of our genome. The HERV transcriptional activity is finely regulated in late developmental stages and the HERV expression is modulated in different cell types and tissues. The consequences of such activity may have an impact on both human physiology and pathology. Anyway, up to date, the HERVs contribution to our biology is only partially understood, often due to the poor characterization of the involved loci. For this reason, the comprehensive identification, classification and characterization of the HERV loci lay the foundations for studies of HERV expression and modulation. Moreover, novel high-throughput sequencing tools have recently allowed a great advancement in elucidating the various HERV expression patterns in different tissues, the control mechanisms of their transcription, and it overall helped in getting better insights in an all-inclusive understanding of the impact of HERVs in the biology of the host.

In this work, we firstly focused on the analysis of the HML-6 group, a member of the class II Betaretrovirus-like. This group includes several proviral loci with an increased transcriptional activity in cancer. One HML-6 locus encodes the small protein ERVK3-1, expressed in various healthy tissues. Moreover, another HML-6 locus encodes HERV-K-MEL, a small Env peptide expressed in samples of cutaneous and ocular melanoma, but not in normal tissues. We characterized the group, reporting the distribution and genetic composition of 66 HML-6 elements. We analyzed the phylogeny of the HML-6 sequences and identified two main clusters. We provided the first description of a Rec domain within the env sequence of 23 HML-6 elements. A Rec domain was also predicted within the ERVK3-1 transcript sequence, revealing its expression in various healthy tissues. We reported the co-localization of 19 HML-6

elements with functional human genes. Indeed, we provided the first complete overview of the HML-6 elements in GRCh37(hg19), describing the structure, phylogeny and genomic context of insertion of each locus.

Secondarily, we used a bioinformatic approach, based on RNA-sequencing (RNA-seq), to study the expression and modulation of HERVs in a scenario of immune system activation. We analyzed a dataset of Human Peripheral Blood Mononuclear Cells (PBMCs) RNA-seq from i) 15 healthy participants before and after the exposure to Lipopolysaccharide (LPS) ii) 19 subjects before and after the administration of an inactivated vaccine. We described the HERV transcriptome in PBMCs, finding that about 8 % of the HERV/MaLR loci were expressed, and identifying the Beta-retrovirus HERVs as those with the highest percentage of expressed loci. We found loci that were modulated as a result of both stimulation with LPS and vaccine administration. The HERV-H group showed the highest number of differentially expressed most intact proviruses. We characterized the HERV loci differentially expressed, checking their genomic context of insertion. In case of the LPS stimulation, that induces a strong activation of innate immune response, we observed a general co-localization with genes that are involved and modulated in the immunity. The analyses showed that HERVs and MaLRs are expressed in PBMCs and regulated in inflammatory settings. The modulation patterns of HERVs and MaLRs are different after LPS stimulation and vaccine administration, presumably indicating that such modulation patterns differ among innate and adaptive immune response.

Publications

The work described in this PhD thesis has been presented in the following manuscripts:

1. *Pisano, M. P.; Grandi, N.; Cadeddu, M.; Blomberg, J.; Tramontano, E. (2019) 'Comprehensive Characterization of the Human Endogenous Retrovirus HERV-K(HML-6) Group: Overview of Structure, Phylogeny, and Contribution to the Human Genome', Journal of Virology, 93(16), pp. 1–19. doi: 10.1128/JVI .00110-19.*
2. *Pisano, M. P.; Grandi, N.; Tramontano, E. (2020) 'High-Throughput Sequencing is a crucial tool to investigate the contribution of Human Endogenous Retroviruses (HERVs) to human biology and development' (Viruses 12, 6; doi: 10.3390/v12060633*
3. *Pisano, M. P.; Tabone, O.; Bodinier, M.; Grandi, N.; Textoris, J.; Mallet, F.; Tramontano, E. 'RNA-seq transcriptome analysis reveals LTR-retrotransposons modulation in Human Peripheral Blood Mononuclear Cells (PBMCs) after in vivo Lipopolysaccharides (LPS) injection' (under review in Journal of Virology)*
4. *Pisano, M. P.; Grandi, N.; Tramontano, E. 'LTR-retrotransposons expression and modulation after inactivated vaccine administration' (Manuscript in preparation)*

1. Introduction

1.1 Retroviruses

The members of the family *Retroviridae* are animal and human pathogen belonging to the group IV of the Baltimore classification, which includes RNA positive-stranded enveloped-viruses with a DNA intermediate in their life-cycle [1,2]. Indeed, after virus attachment and penetration inside the host cell, retroviruses employ two viral enzymes, the Reverse Transcriptase (RT) and the Integrase (IN), to reverse transcribe their single-stranded RNA genome into a double-stranded DNA, and to integrate this DNA into the genome of the host [3]. The integrated form of the viral genome is named provirus, and the proviral genes can be transcribed through the host translational machinery [3]. The current taxonomy of retroviruses is based on the 2019 release of the International Committee on Taxonomy of Viruses (ICTV) [2]. According to this taxonomy, the Retroviridae family can be divided in two subfamilies, Orthoretrovirinae and Spumaretrovirinae that include, respectively, 6 and 5 different genera.

The genome of the proviruses can be simple or complex. Simple proviral genomes present four genes *gag*, *pro*, *pol* and *env*, flanked by two Long Terminal Repeats (LTRs). The *gag* gene codes for the proteins Matrix (MA), Capsid (CA) and Nucleocapsid (NC); the *pro* gene for the protein Protease (PR); the *pol* gene for the proteins Reverse Transcriptase (RT), Ribonuclease H (RH) and Integrase (IN); and the *env* gene for the proteins Surface (SU) and Transmembrane (TM). The complex genomes have a similar structure but include some additional genomic information for coding small accessory proteins with different functions, for example the Rex protein of the Human T-cell Leukemia Virus (HTLV) or the Rev protein of the Human Immunodeficiency Virus (HIV). The LTRs are constituted by a Unique 3 (U3) sequence, a short Repeated (R) sequence and a Unique 5 (U5) sequence. The U5 and U3 regions contain regulatory sequences needed for both the provirus integration and the regulation of the viral gene expression (e.g. promoters, enhancers and poly-A signals) [4].

The first step of the lifecycle of retroviruses is the attachment of their glycoproteins SU and TM to specific receptors on the membrane of the target cell. This interaction allows the entrance of the virus into the cytoplasm, inducing the fusion of the envelope with the membrane. After that, RT mediates the retrotranscription of the RNA into a double-stranded DNA. The viral DNA is hence translocated to the nucleus, where IN integrates it within the genome of the host. After the integration, the genes of the provirus are recognized as cellular genes and transcribed by the cellular RNA polymerase. The viral mRNA is translated producing proteins and polyproteins, which are then cleaved into functional subunits and used to assemble new virions. The new virions spread from the cell by exocytosis, and the envelope is produced from the cell membrane (Figure 1) [4,5].

Generally, retroviruses infect somatic cells and can be horizontally transmitted within members of a host population. However, some retroviruses can also infect cells of the germline, and occasionally the integrated proviral genomes can be vertically transmitted to the offspring. Therefore, the provirus can be transmitted among the generations and eventually fixed within the population [4].

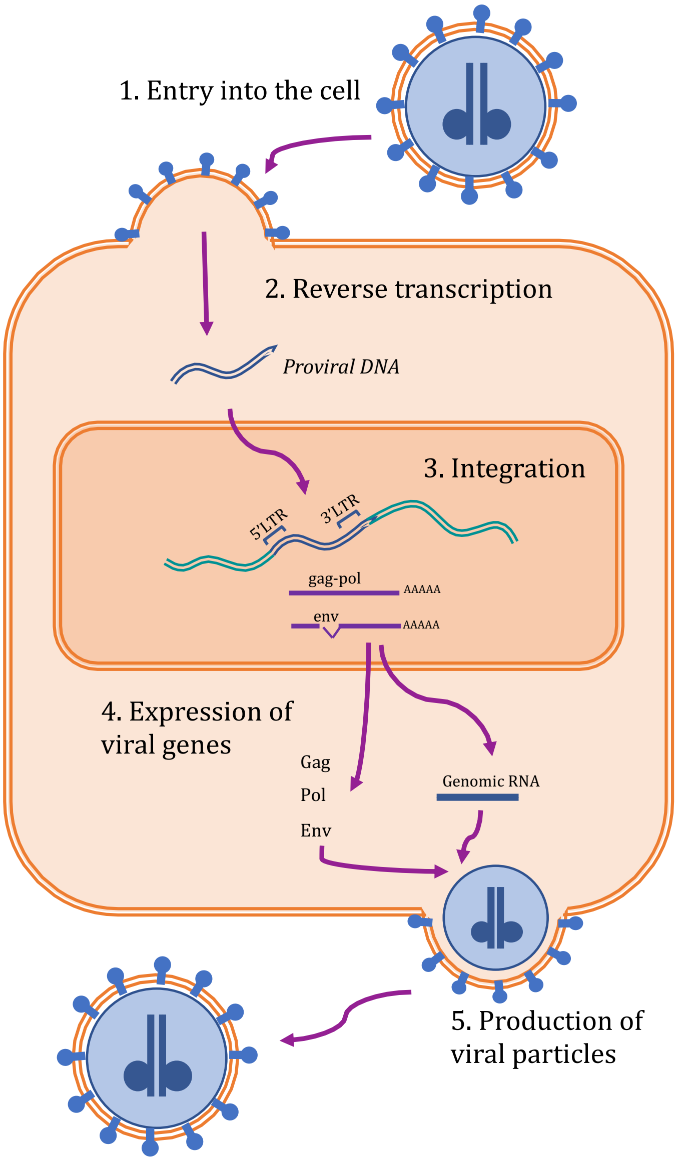


Figure 1. Schematic overview of the retrovirus life-cycle. The virus enters into the cell (1); its RNA genome is reverse transcribed into ds-DNA (2); the viral DNA is integrated into the host DNA (3); the viral genes are expressed (4) and new viral particles are produced and leave the cell (5).

1.2 Human Endogenous Retroviruses (HERVs)

A large proportion of the human genome consists of repeated elements, including Human Endogenous Retroviruses (HERVs). These elements are remnants of ancestral and independent retroviral infections within the germline cells that took place millions of years ago [4] (Figure 2). At the time of integration, the HERV genome was composed of the four retroviral genes flanked by the two LTRs. Some more ancient retroviral elements, the Mammalian apparent HERVs and MaLRs (MaLRs), had a similar genomic structure but lacking the *env* gene [6,7]. Over time, most of these elements have accumulated abundant mutations, often compromising their coding capability. A great number of HERV insertions are now present as solitary LTRs, generated by recombination occurrences [7], as clearly observably when comparing human HERV integrations with their orthologs in primates [7,8].

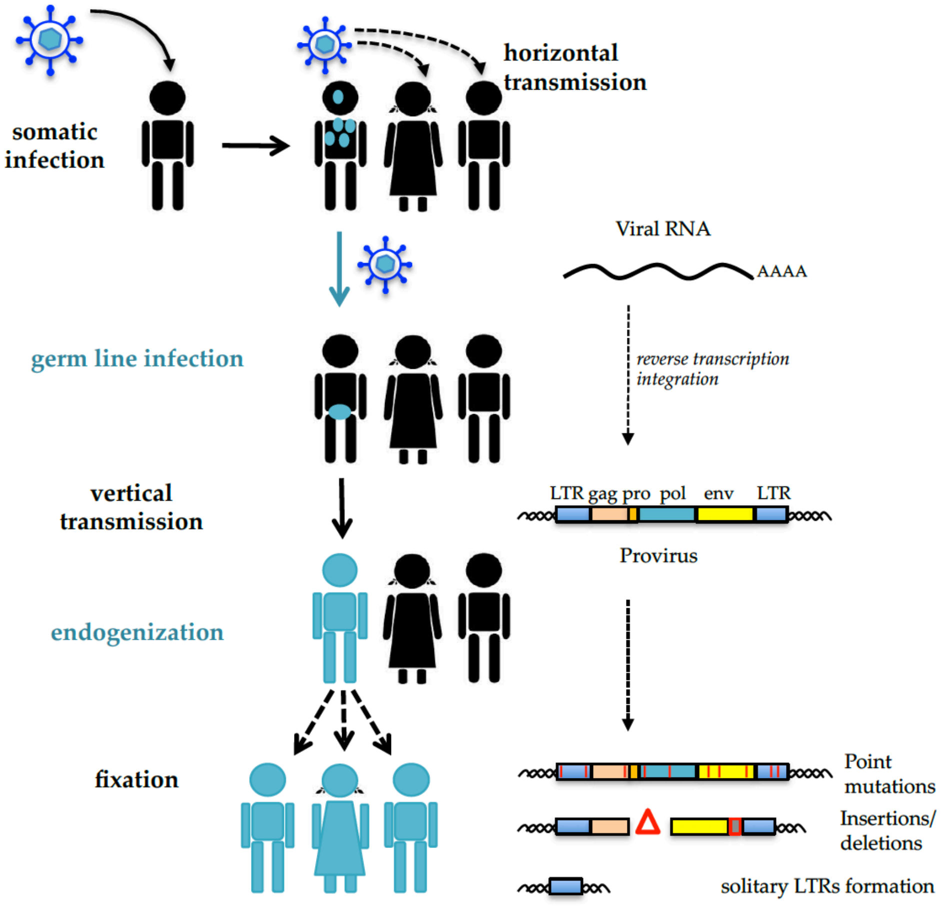


Figure 2. Retroviruses endogenization and HERVs formation. Retroviruses usually target the somatic cells, showing a horizontal transmission from an infected individual to new hosts. Some retroviruses can infect the germ line cells, which are transmitted to the offspring. The vertical transmission of HERVs has determined, over time, their fixation into the human genome. During evolution, the majority of HERVs accumulated multiple mutations that generally compromised their coding capacity, often causing the elimination of the internal portion, leaving solitary LTRs. From Grandi *et al.* 2018.

1.2.1 HERV identification and classification

For a long time the identification of HERVs has been a bioinformatics challenge [9]. An important tool for the identification of HERVs in the human genome is RepeatMasker (http://www.repeatmasker.org), a program that checks the genomes for interspersed repeats, by making use of a database of HERV references, Repbase (https://www.girinst.org). RepeatMasker also makes use of Dfam (https://www.dfam.org), another database of repetitive sequences organized by families. Of note, the analysis of RepeatMasker allows collecting the majority of repetitive elements referred to HERVs and solo LTRs, but it is not able to predict the retroviral structure of HERV proviruses. Another HERV database, hervgdb4, has been created with the specific aim to detect HERVs though an Affymetrix array (HERV-V3). Hervgdb4 includes proviral and solo LTR sequences that have been collected by using 42 selected proviral sequences (prototypes) as references for RepeatMasker analyses or, alternatively, by reconstructing proviral structures from data of the Dfam database [10]. Of note, the "prototype" subset of sequences into the hervgdb4 database also includes gene annotation. Since this database has been created to design the probes of an Affymetrix array, all the sequences are fragmented [10].

Recently, it has been developed a tool, named RetroTector, for the automated recognition of the best-preserved proviral sequences in the genome of vertebrates [11,12]. The analysis of RetroTector identified a total of 3173 best-preserved HERV sequences in the human genome assembly GRCh37 (hg19) [7]. Importantly, the collection of the best-preserved HERVs has allowed further phylogenetic studies for the classification and characterization of all the HERVs and HERV groups [7]. Indeed, also the HERV classification has been challenging for a long time. This classification work split HERVs in three main classes: class I (Gamma-retrovirus- and Epsilon-retrovirus-like), class II (Beta-retrovirus-like) and class III (Spuma-retrovirus-like) [7]. Similar elements have also been clustered into 70 phylogenetic groups, of which 39 "canonical" and 31 "non-canonical" clades characterized by several degrees of mosaicism (Table 1) [7]. RetroTector also predict the sequence of the retroviral genes and a multitude of other retroviral features, like the Primer Binding Site (PBS) or the Poly Purine Tract (PPT). By contrast, it is unable to detect solo LTRs [11]. Finally, the analysis of RetroTector revealed that Mammalian LTR Retrotransposons (MaLRs) are the most common retroviral components in the human genome. MaLRs are ancient LTR-retrotransposons characterized by the lack of the *env* gene, which, anyway, have not been deeply characterized and classified in that work [11].

Table 1 General HERV identification and preliminary classification in GRCh37/hg19 (From *Vargiu et al 2017*)

Probable genus	Type species	HERV genus	Nr of total sequences	Nr of clades

Gammaretrovirus and Epsilonretrovirus	Murine leukemia virus (MLV)
	Feline leukemia virus (FeLV)
	Walleye dermal sarcoma virus (WDSV)
		Class I (gamma-like, epsilon-like)	2341	Canonical 27;
				Noncanonical 25;
				Total 52

Betaretrovirus	Mouse mammary tumor virus (MMTV)
	Mason-Pfizer monkey virus (MPMV)
	Jaagsiekte sheep retrovirus (JSRV)
		Class II (beta-like)	598	Canonical 10;
				Noncanonical 0;
				Total 10

Spumaretrovirus	Simian foamy virus (SFV)	Class III (spuma-like), including MaLR (i.e. MST-MLT-THE)	216	Canonical 2;
				Noncanonical 5;
				Total 7

Errantivirus	Gypsy retrovirus	Uncertain_Errantilike	2	Canonical 0;
				Noncanonical 1;
				Total 1

		Unclassifiable	16	
		Total	3173	

In addition to the mentioned tools for HERV identification and reports on their classification, there are also some studies that collect and characterize the HERV sequences belonging to a single group. This kind of study starts with the analyses of the human genome with RepeatMasker, RetroTector, or both. The data collected are then manually visualized and inspected, and sometimes implemented by performing BLAT searches. The HERV coordinates provided in these works are hence the most accurate and well-annotated. Example of HERV groups deeply studied are the HERW, and several HML subgroups [13–18]. Moreover, the characterization of the HML-6 group is also one of the aims of this thesis.

1.2.2 HERV impact on human biology

An overall few HERV insertions have been studied to understand their implication in human physiopathology [19]. The best-described example of HERV involvement in the host physiology is the production of syncytin-1, a retroviral protein coded by the env gene of a provirus belonged to the HERV-W group, expressed in trophoblasts. This protein has an essential role in driving the placental mammals' evolution, as it is necessary for placental development [20,21]. Differently, other HERV proviruses and proteins have been investigated for their possible involvement in pathogenesis, in particular proposing a possible role in both cancer and autoimmunity [5]. In fact, there are several pieces of evidence showing an abnormal increase in HERV expression in tumor cells [22] and HERV Env proteins may provoke cell fusion and may thus potentially promote tumor development [23,24]. Moreover, two accessory proteins, termed Np9 and Rec, produced from splicing variants of the env gene in proviruses belonging to the HML-2 group, may have oncogenic properties [25,26]. The complex connection between HERVs and the immune response has been also widely investigated [27,28]. Indeed, some inflammatory settings can induce HERV expression, while some HERV products may trigger the host immune response and hence activate the innate immune pathways [27,28]. For instance, one provirus from the HERV-W group encodes an Env protein potentially associated with multiple sclerosis [29,30]. This protein has been shown to induce inflammatory effects, possibly acting as a superantigen, and a monoclonal antibody recognizing it is currently under clinical trial [31]. Importantly, extensive knowledge of the mechanisms of HERV-mediated immune activation would be essential for the understanding of possible HERV implications in inflammatory conditions as well as in autoimmune diseases [28].

HERVs can also have an impact on human biology other than through proteins production [19,32]. HERV LTRs include enhancers, promoters, polyadenylation signals and splice sites within their sequences and may influence neighboring cellular gene expression [33,34]. One of the most significant examples of HERV-controlled human gene expression is the HERV-E LTR integrated upstream of the pancreatic amylase gene that acts promoting its expression [35]. In addition, HERV integrations may alter the normal gene functions by providing alternative and aberrant sites for splicing or by interfering, either positively or negatively, with the mRNA transcription through the production of non-coding RNAs [28,36].

Given that HERVs/LTRs represent a large portion of the human genome, and can potentially influence our physiopathology, it is quite clear that cells should finely control the HERV transcriptional and translational activity through various mechanisms such as accumulation of mutations, RNA silencing, or histone/DNA methylation [15,37–40]. While most HERVs are silenced, some elements are normally expressed in various developmental stages of human embryogenesis, and their activity is regulated in different human tissues [40,41]. Indeed, HERVs may be involved in creating Topologically Associating Domains (TADs) during pluripotent stem cell differentiation, thus helping to define a three-dimensional organization of chromatin facilitating interactions between enhancers and promoters [42]. HERVs could be also activated as a consequence of some pathological conditions, like HIV infections or cancer, characterized by alterations in epigenetic regulation [43–46]. Overall, such expression patterns make difficult to clearly establishing a causal association between HERVs and diseases.

Recently, it has been shown that some recently integrated members of the HML-2 group determined the presence in the human population of insertional polymorphic integrations [47], that might be important to fully understand the role of HERVs in human biology. Indeed, the identification of such polymorphic HERVs integrated into regions of the genome that are essential for phenotypic functions allowed to predict HERV insertions co-occurring with known pathological variants [48].

1.2.2.1 HERV variability in human population

The HERV-K (HML-2) group is known to include some young elements integrated into the genome of modern humans after the divergence from the lineage of chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*) [13]. Moreover, the HML-2 group may have been active in archaic hominids also after the divergence from the lineage of modern humans [49,50]. Indeed, some studies have analyzed the genome sequences of Neandertal and Denisovan, identifying 14 HML-2 insertions not included in the modern human genome assembly [49,50]. Only some of these insertions in archaic hominids were also found into the genome of certain individuals of modern humans, as unfixed loci [49,51,52]. Hence, such recent insertional activity of the HML-2 group is of interest for the consequent possible presence of polymorphic proviruses [13,47]. In total, there are 36 HML-2 proviruses in human population that are not included in the human reference genome [48–53]. Furthermore, also some of the proviruses present in the human genome assemblies hg19 and hg38 are known to be unfixed among the population [13,53,54]. The analysis of data from the 1000 Genomes Project has revealed differences in the frequency of HML-2 insertions among the five super-populations: African, East Asian, Ad Mixed American, European, and South Asian. Not all the HML-2 insertions occur in the different populations with the same frequency

[52,53], and the state of presence or absence of the totality of the proviruses is sufficient to distinguish the five super-populations [53]. Many of the unfixed insertions are rare, and the East Asian population is the one with the lowest prevalence of HML-2 insertional polymorphisms [52,53]. About the potential role of these loci, interesting data derive from the insertions that have significant Single-Nucleotide Polymorphism (SNP) association enriched for Expression quantitative trait loci (eQTLs). Indeed, such information tries to establish a relationship between a single nucleotide variant for HML-2 polymorphic and tissue-specific gene expression [48,55]. Interestingly, 46 insertional polymorphisms have SNPs enriched for eQTLs across 44 human tissues [48]. Moreover, 15 of them have SNPs associated with specific neurologic and immunologic traits, including Parkinson's disease and other autoimmune diseases [48].

Sequence variances in HERVs that are not polymorphic insertions may also be very informative. For example, polymorphisms occurring in transcription factor binding sites may explain the differential HERV expression among individuals and in cancer [56]. Moreover, many HERV elements have been found to be enriched for somatic mutations (hotspots) in cancer [55]. Among these hotspots, the mutation C2270G in ZNF99 is associated with a lower survival rate in kidney cancer patients, and it can be potentially used as a biomarker [55].

1.2.2.2 HERV expression is regulated during human development

The control of HERVs and solo LTRs in somatic cells has to be tight and well-structured. The HERVs are regularly expressed in the germline, but the epigenetic regulation is essential to finely control their expression since the first steps of embryogenesis [40,41,57]. It is well-known that the DNA of embryonic stem cells is hypo-methylated, leading to a general release of HERVs [57]. In particular, the HML-2 and HERVH groups are extremely transcriptionally active and HERVH elements are also directly involved in the maintenance of pluripotency [58–60].

The HML-2 group has been reported to be expressed during the early embryonic stages, in cells from morula and pre-implantation blastocysts [60]. The proviral RNAs result in the production of retroviral products, such as Gag proteins and the HML-2 accessory protein Rec [60]. Through the binding of HERV RNAs, Rec promotes the ribosomal targeting and hence the HERV RNAs translations. Moreover, this protein may potentially provide a protective effect against viral infections by inducing restriction pathways [60].

Naïve embryonic stem cells are derived from blastocysts. Importantly, as the human naïve cells are not easy to be isolated, human naïve-like cells are artificially generated to be functionally equivalent to those from blastocyst-stage embryos [61,62]. The expression pattern of HERVH elements is a key component of pluripotency, and a useful application is to use their expression as a marker for capturing the human naïve pluripotent state *in vitro* [63,64]. HERVH elements provide functional binding sites for transcription factors driving the production of chimeric transcripts that modulate pluripotency acting as long ncRNAs [61]. About 639 of these LTR-associated RNAs have been found in human embryonic cells [58]. An example of a transcription factor that binds a HERVH LTR is the LTR-binding protein 9 (LBP9) [61,65]. The chimeric transcript produced after the LBP9-HERVH binding is necessary for the maintenance of naively: the absence of LBP9, HERVH, and HERVH alternative transcript drives to the loss of the pluripotency state of the cells [61,65]. Similarly, another HERVH LTR includes binding sites for NANOG protein [63], and also the recruitment of the Octamer-Binding Transcription Factor 4 (OCT4) to the binding site of a HERVH LTR7 drives to the expression of chimeric pluripotency-associated transcripts [59].

Among the HERVH group, not only the LTRs are active to produce chimeric long ncRNAs, but also several proviruses are transcribed. The classical RNA structure from expressed proviruses is 5'LTR-gag-pro-3'LTR, but their sequences seem not to include intact open reading frames. In general, HERVH RNA constitutes about 2% of all poly-A RNA in embryonic cells [63].

While in naïve cells the HERVH elements contribute to maintaining the pluripotency, during pluripotent stem cell differentiation this group is involved in shaping species-specific chromatin architecture [42]. Indeed, HERVH elements actively contribute to the creation of transcriptionally active and self-interacting compartments, TADs [42]. The HERVH elements act creating TAD boundaries, and such ability is dependent on transcription, probably due to the positioning of coesin in complexes mediated by polymerase II movements [42,66]. Importantly, the creation of HERV-mediated TAD boundaries suggests that these elements can have an important impact on gene regulation. Hence, the spreading of HERVs has contributed to the evolution of the chromatin architecture [67,68].

The repression of HERVs expression is established in the pre-implantation embryo and then maintained in most developed tissues [69]. The mechanisms of HERVs silencing are various, and many data refer to murine Endogenous Retroviruses (mERVs), in a mouse model. Firstly, DNA methylation, catalyzed in mice by DNA methyltransferases, is required to repress mERVs in differentiated cells [69,70]. Anyway, DNA methylation is not the only mechanism controlling mERVs expression during embryogenesis. For example, the protein Histone-lysine N-methyltransferase SETDB1 may have a critical role in inhibiting mERVs expression, as it is evident in SETDB1 knockout embryonic cells, where several mERVs are derepressed [71]. Indeed, HERVs may take advantage of the Ten-eleven translocation methylcytosine dioxygenase (TET) class of proteins to evade DNA methylation-mediated transcriptional repression. For this reason, cells may have evolved methylation-independent silencing pathways, like histone modification, during developmental stages or tissues when DNA methylation is compromised [70]. KRAB zinc finger proteins (KZFPs) are also involved in silencing HERVs through targeting repressive chromatin states [72,73]. Indeed, KZFPs and their cofactor Tripartite motif-containing 28 (TRIM28) promote chromatin modifications to regulate HERV transcriptional activity [73]. Finally, a microRNA (miR-34a) can repress mERVs expression by restraining some transcription factor binding proteins [74].

1.2.2.3 HERVs contribute to somatic cells physiology and disease

HERVs may be transcribed also in somatic cells [67,75]. A study has analyzed the HERV expression in RNA-Seq samples from the ENCODE project, finding HERVs active in a cell line-specific manner [76,77]. Aging does not have a strong effect on the overall HERV expression, but several proviruses are moderately affected and it seems that there are some age-dependent expression patterns [78]. Of course, as the consequences of HERVs activity in somatic cells may be deleterious, the host makes a great effort to efficiently repress the great majority of HERV expression [79,80]. Indeed, 794,972 LTR sequences have Transcription Factor-Binding Sites (TF–Ss) - most of which co-localized with genes involved in the immune response - that may potentially interfere with neighboring genes [75]. The activity of HERV proviruses and solo LTRs is modulated in response to stress and immune activation [45,81,82]. Importantly, such a modulated response is different from the one observed in cancer. The expression of certain HERV loci induced by stress and immunity is different from the one occurring in cell transformation, in which there is a widespread expression of these elements [45]. For example, among most LTRs including interferon-inducible enhancers [83], some MER41 elements include STAT1- and IRF1-binding sites and mediate the activation of the response to pathogens [67,75]. The high number of regulatory elements linked to the immune response is not accidental [83]. In fact, by introducing and amplifying interferon-sensitive enhancers, HERV integrations have shaped the evolution of transcriptional pathways that define the interferon response [83]. An interesting mechanism of HERV activation in immunity is based on the loss of Tripartite Motif-containing 28 (TRIM28), triggered by the influenza virus. TRIM28 usually silences the HERV expression, but the viral infection modifies the status of TRIM28 and alleviates the HERV repression. The consequence is a production of double-stranded (ds)RNA that triggers the dsRNA-activated IFN-mediated defense [81]. Indeed, not always the HERV modulation results in an increase of HERV expression, but also in a decrease in expression, as in the case of the hippocampus. This area of the brain is particularly susceptible to stress, and here the acute stress is correlated with the silencing of HERVs [82]. HERVs can be also modulated by the histone deacetylases inhibitor vorinostat, which reactivates HIV in latently infected cells. Overall, about 2,000 HERV loci are significantly modulated by vorinostat, several HERVL elements were predominantly downregulated, in contrast to HERV-9 elements that were mostly upregulated [84].

HERVs are known to be de-silenced and hence transcriptionally active in cancer. Indeed, the HERV transcriptional activity has been found significantly higher in cancer cells than in controls [77]. A mechanism of HERV modulation is mediated by lysine acetyltransferase TIP60. TIP60 is a tumor suppressor that silences retro-transposon elements. In cancer, this protein is downregulated and the loss of its activity results in HERV de-repression [85]. In colon cancer, the activation of six LTR promoters affects the expression of cellular genes [86]. Moreover, in a subset of colon cancer samples, an LTR promoter is coded in a chimeric transcript involving the Interleukin 33 (IL-33) gene, which encodes an aberrant isoform of the protein [86]. HERV activity is also increased in human breast cancer, in tumorigenic cell lines. In particular, the HML-2 proviruses in loci 3q12.3 and 11p15.4 display increased activity in almost all the tumorigenic breast cell lines [56]. Interestingly, the increment of the HML-2 expression is higher in breast basal-epithelial then in other cells. Moreover, in these cells, the *env* gene is the most upregulated [87]. For these reasons, the HML-2 elements are possible biomarkers for this particular form of breast cancer, or they can be a target for cancer vaccines or immunotherapy [87]. The reactivation of LTR promoters influences the expression of neighboring genes also in renal cell carcinoma. In particular, several LTRs with hypoxia-inducible transcription factors are activated. For example, an LTR promoter upstream of the stem cell transcription factor POU5F induces the production of an aberrant transcript POU5F1 isoform [88].

Patterns of HERV modulation are evident also in other diseases. For instance, some HERV-W and HERVH loci have been found to be expressed in postmortem brain samples from schizophrenia and bipolar patients [89]. The HERV-W group has been often tentatively correlated to multiple sclerosis [29,30]. Hence, to investigate the expression and modulation of HERV-W loci in multiple sclerosis, it has been analyzed the HERV-W transcriptome in brain lesions [90]. Interestingly, transcript levels of HERV-W loci were similar in healthy samples and multiple sclerosis, suggesting a lack of HERV-W modulation correlated to the disease [90]. Analyses of HERV expression have led to contrasting results in PBMCs from patients with systemic lupus erythematosus. A first work observed a general trend of HERV downregulation [91], while a second one identified 124 significantly upregulated HERVs, and none downregulated [92]. HERVs are also associated with drug addiction. A polymorphic

HML-2 solo LTR is an antisense integration within the sequence of a gene that affects dopaminergic activity [93]. The expression of this antisense LTR can modulate the expression of the neighbor gene, and integration is more frequently present in drug-addicted then in the general population [93].

1.2.3.4 Beta-retroviruses and the HML-6 group

Between the HERV groups, the Human MMTV-like (HML) supergroup of class II is one of the most investigated, mainly due to the fact it includes some of the youngest and best conserved elements, belonging to the HML-2 clade [13]. This supergroup consists of 10 clades (HML-1 to -10) that are related to the exogenous Mouse Mammary Tumour Virus (MMTV) [7]. It has been reported that some HML-2 elements are able to encode for an mRNA nuclear export protein, Rec, coded from a doubly spliced transcript which is a functional homolog of the retroviral regulatory proteins MMTV Rem [94], HIV Rev and HTLV Rex [95,96]. The HML-2 rec accessory gene can be present in two forms, a first one with full-length sequence (characteristics of type 2 HML2 elements), and a second one with a 292-bp deletion that codes for a smaller protein name NP9 (associated with type 1 HML2 sequences) (17, 18). Recently, the Rec domain has also been found within the genome of some HML-10 elements [17].

In addition to HML-2 and HML-10, HML-6 is also a highly investigated HML clade. The earliest studies about this group collected 10 sequences identified by using a PCR approach with HML-6 specific primers [98,99]. A first phylogenetic characterization of these elements allowed describing the HML-6 subgroup as a heterogeneous but distinct group of elements belonging to the HERV-K superfamily, with a PBS for lysine tRNA [98,99]. The HML-6 Betaretroviral features were also detected: two Zinc-fingers in *gag*, and both dUTPase and G-Patch domains in *pro* [98,99]. Notably, the dUTPase tree showed a different phylogeny from the one of the other genes, and further analysis concerning the presence of dUTPase in the various HERV-K subgroups demonstrated that the HML-6 dUTPase sequences appear to be more related to the MMTV dUTPases than to those of the other HML members [98,100]. According to RetroTector analysis, this subgroup includes 48 canonical elements and additional 17 non-canonical elements coming from recombination events [7], and the internal sequences are flanked by two LTRs identified among four types (LTR3, LTR3A, LTR3B and LTR3B_v) by RepBase.

A first important study reported an extensive transcriptional activity of HML-6 elements through retrovirus-specific microarray [101]. HML-6 transcripts were found in all the 19 healthy tissues analyzed [101]. Of note, an HML-6 element in locus 19q13.43b that contains an intact open reading frame (ORF) was reported to encode a small transcript, ERVK3-1, expressed in various healthy tissues (ENSG00000142396), and gave support to the hypothesis of an extensive HML-6 expression activity [101]. Subsequently, besides physiological expression, HML-6 sequences were reported to be of particular interest due to either the selective activation, or the increased activity, of several proviral loci in malignant mammary gland tissue from patients with human breast cancer [102]. Other examples of HML-6 expression in cancer were also reported in cutaneous and ocular melanoma cells, in which a small peptide from an HML-6 env gene, namely HERV-K-MEL, was detected in tumor tissues but not in normal tissues [103]. Despite these findings prompted the possibility of a HML-6 contribution to diseases, the causal relationship between HML-6 expression and cancer is still not clear, and further expression studies of the individual HML-6 loci are needed to clarify their potential contribution to human pathogenesis.

1.2.3.5 HERV contribution to the immune response

The HERV contribution in shaping and influencing the human innate immunity is an argument of particular interest [28]. Indeed, in some cases, HERV derived antigens could be recognized as pathogen-associated molecular patterns (PAMPs) or Danger-Associated Molecular Patterns (DAMPs), by Pattern Recognition Receptors (PRRs) such as the transmembrane Toll-Like Receptor proteins (TLRs) [27,104] (Figure 3). In these cases, the activation of PRRs evokes complex cellular signaling pathways altering gene expression to transduce pro-inflammatory signals [28]. Even though on the one hand it has been hypothesized that these interactions with PRRs contributed positively in shaping the evolution of the immune response [27,105], on the other hand the same mechanisms have been investigated for their possible contribution to the development of autoimmunity and inflammatory diseases [5,106–108], like multiple sclerosis [109,110]. Accordingly, the activation of the immune response through treatments with LPS or TNF-α can lead to an increase of HERV expression [111,112]. For instance, a recent microarray-based study revealed the *in vitro* modulation of HERV expression in PBMCs after high- and low-dose LPS and Interferon-γ (IFN-γ) stimulation [111]. A similar approach allowed to observe HERV *in vivo* modulation in samples of blood in various contexts of injuries, also introducing a possible role of HERVs close to immunity-related genes in the regulation of their expression [46]. In any case, many questions on the actual role of HERV expression in immunity are still unsolved, and - in this respect - the characterization of individual HERVs' genomic localization and coding capacity could help to understand their potential effects [32].

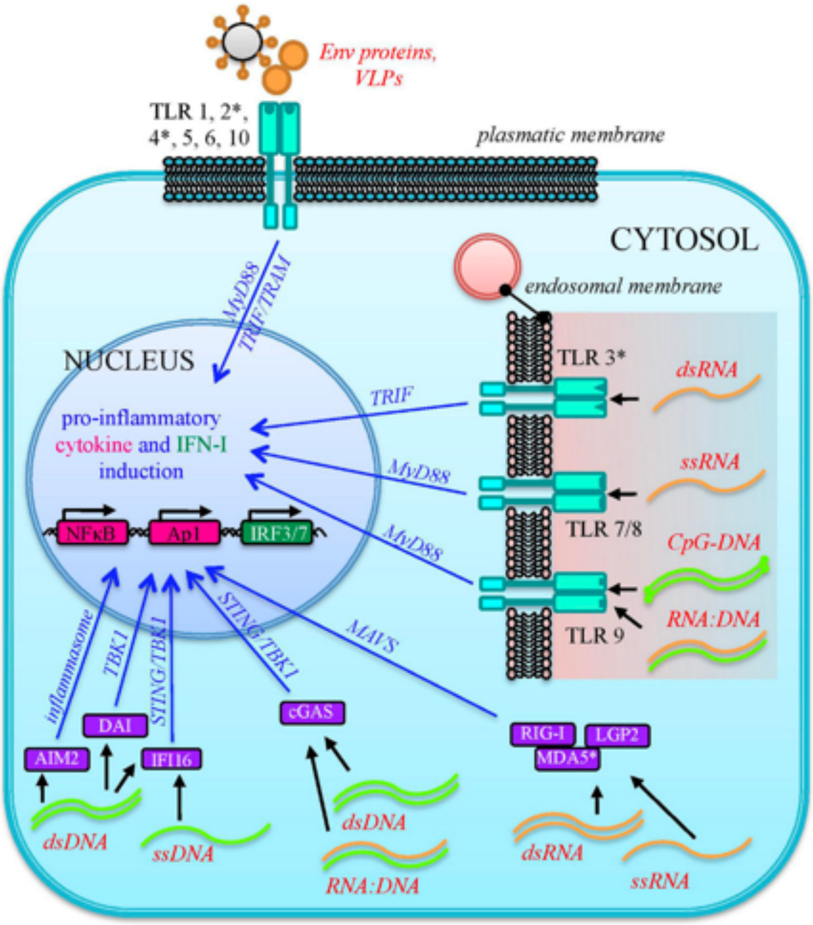


Figure 3. Sensing of HERV molecules by innate immunity PRRs. HERV proteins and nucleic acids can be recognized as PAMPs or DAMPs by PRRs or Toll Like Receptors (The star highlights specific receptor that can interact with HERVs). The link with PRRs triggers signaling cascade for the activation of immune genes encoding for pro-inflammatory effectors, like cytokines and type I IFN.

1.3 Bioinformatic and high-throughput applications in HERV research

The great majority of the studies on the effects of HERVs on human pathophysiology are based on microarrays, hybridization-based approaches, or Reverse Transcriptase Polymerase Chain Reaction (RT-PCR). Unfortunately, due to technical limitations, these studies have often failed to explain the complexity of the HERV impact on host biology in its entirety [113]. However, the sequencing of the human genome, the resulting information of the genomic characterization of HERVs and, finally, the advent of high-throughput technologies have led to a great advancement in this field [114]. In fact, such technologies have allowed to take into account genome variations, to analyze regulatory elements and three-dimensional organization of the genome, and to characterize the HERV transcriptome [115,116].

Firstly, high-throughput sequencing technologies have allowed performing multiple genomes and transcriptome sequencing in parallel. For example, DNA sequencing and RNA sequencing (RNA-seq) can help to evaluate human genomic diversity, through the identification of variants and mutations [115]. Also, DNA-protein interactions such as Chromatin Immunoprecipitation sequencing (ChIP-seq) and Methylation sequencing (Methyl-seq) are useful to explain epigenetic changes [116]. Finally, at the transcriptomic level, RNA-seq can be used to analyze the transcriptome and identify modulated genes, while Ribosome sequencing (Ribo-seq) can determine mRNA transcripts that are being translated [116].

1.3.1 Genome sequencing and HERV variability

As already mentioned, there is a certain HERV variability in the human genome. A deep knowledge of such variability may effectively help to better understand the real impact of HERVs in human biology. For this reason, one of the most essential challenge in HERV research is the identification of insertional polymorphisms. For this purpose, several studies take advantage of data from large datasets of whole-genome sequences. Indeed, many projects have provided numerous copies of whole-genome sequences in the form of short DNA fragments, called reads, collected in databases. Whole-genome sequences sources can include different type of data, e.g. data from healthy donors, like those collected by the 1000 Genomes Project [117,118], or from patients, like those from cancer patients collected by The Cancer Genome Atlas (TCGA) [119] and the International Cancer Genome Consortium (ICGC) [120,121].

Reads from these sources can be mapped to a control human genome assembly, and variation from the reference sequence may be useful to identify polymorphic insertions or single nucleotide changes [115,116]. Hence, some bioinformatics tools have been developed to discover insertional polymorphic transposable elements, including one specifically-designed for HERVs [122]. In this way, it is possible to detect insertional polymorphic HERVs and to study their frequency in diverse human populations [51–53] (Figure 4a).

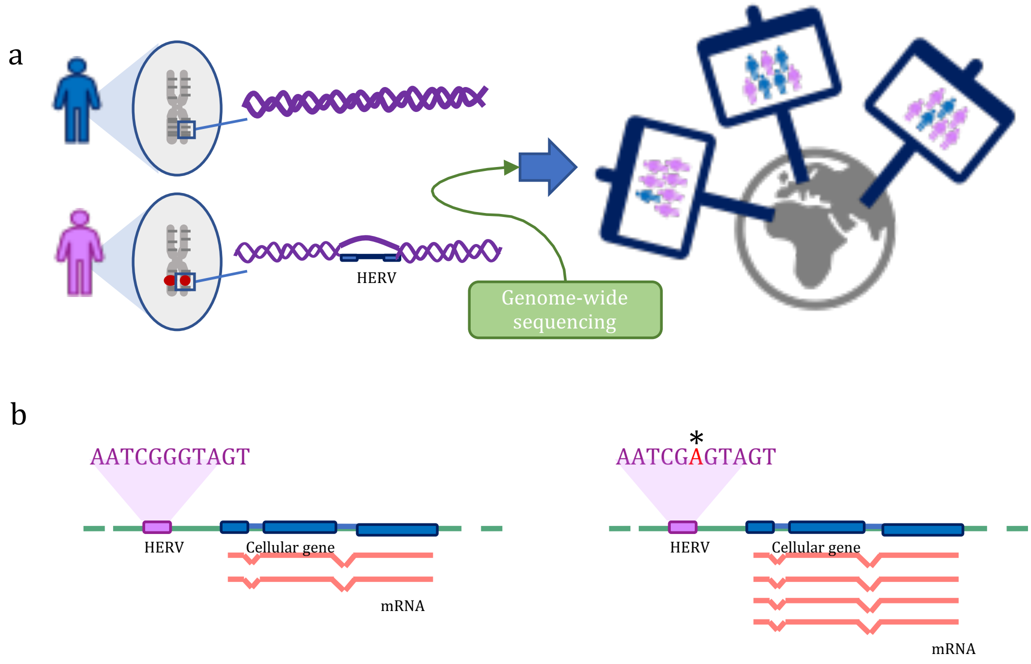


Figure 4. Example of possible application of High-Throughput sequencing on the study of HERVs variability. The High-Throughput sequencing of whole genomes allows the identification of insertional polymorphisms, and to assess the genome-wide distribution of these polymorphic loci (a). This application also allows the identification of single nucleotide variations associated with expression quantitative trait loci (*), which explain changes in the host gene expression levels (b**).**

However, it is still difficult to study insertional polymorphisms within the same individual by using whole-genome sequences data. The application of these technologies remains hence insufficient to investigate HERV mobility or somatic integrations [114]. The identification of SNPs and Single Nucleotide Variants (SNVs) into the HERV sequence is another application that required whole-genome sequences data, especially for its possible implication in diseases [117,118]. Indeed, SNPs may be associated with eQTLs (Figure 4b) [48,55,123]. Such variations in gene expression can be accountable for phenotypical or pathological traits, and HERVs including SNPs associated with eQTLs can hence be directly linked to physiology and diseases [48,55].

1.3.2 Regulation of HERV expression and their involvement in the human gene expression

High throughput sequencing technologies have profoundly improved our knowledge of the three-dimensional organization of the genome, chromatin state, and chromatin modifications [115]. 3D chromatin can be analyzed by paired-end tag sequencing (ChIA-PET) and Hi-C [124], which consist of consecutive steps of legation and sequencing of close cross-linked chromatin portions, and frequently interacting portions can be visualized in contact matrix [115,124]. These technics are particularly useful to identify regions on the same chromosome mostly interacting with each other due to the particular organization of the chromatin in TADs. The main application of these technics in HERV research is the investigation of possible involvements of HERVs in TADs formations (Figure 5a) [67]. The chromatin state and modification can also be studied through the application of ChIP-Seq technologies. Indeed, ChIP-Seq allows analyzing DNA-proteins interactions, through the combination of chromatin immunoprecipitation and high throughput sequencing [115,116]. When applied to the HERV research, ChIP-Seq can be used to check the chromatin states of HERV loci, and their modification in different developmental stages or diseases (Figure 5b) [63,77].

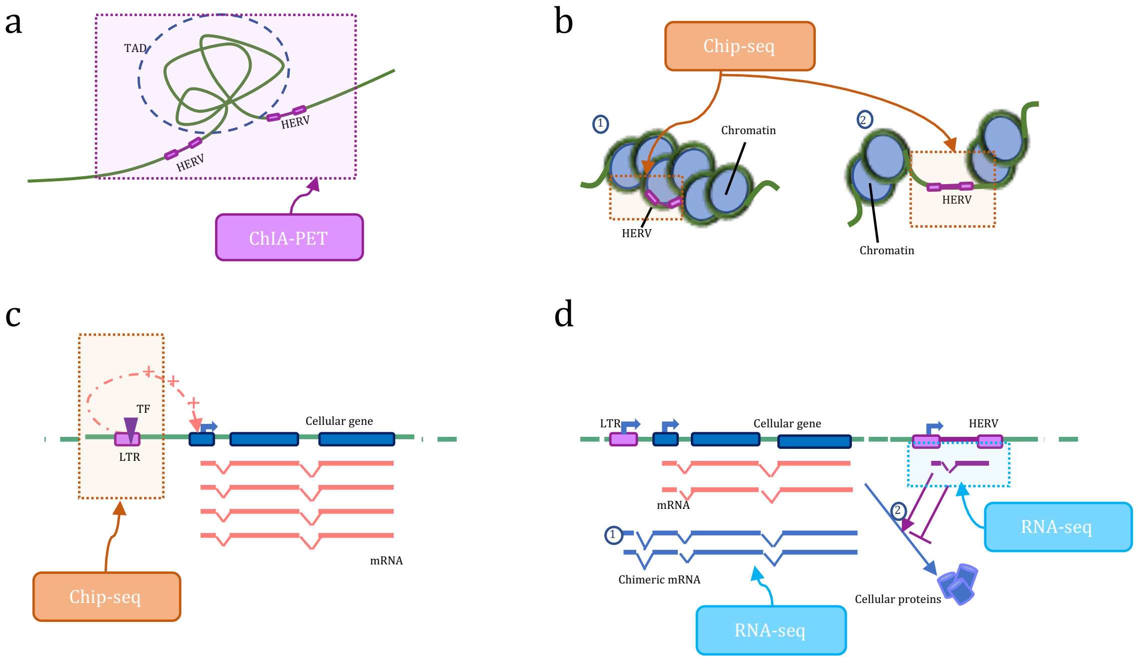


Figure 5. Example of possible application of High-Throughput sequencing to the study of the regulation of gene expression and their involvement in gene expression. Paired-end tag sequencing (ChIA-PET) can be used to analyze 3D interactions of the chromatin. It is possible to investigate HERV involvement in the formation of Topologically Associating Domains TADs (a). Chromatin Immunoprecipitation sequencing (ChIP-Seq) allows to analyze DNA-protein interactions. ChIP-Seq can be used to check the chromatin state of HERV loci (b), which may be transcriptionally silenced (1) or transcriptionally active (2). ChIP-Seq can be also used to study the interaction of DNA with other proteins, e.g. Transcription Factors (TFs). TF-LTRs interactions can enhance the expression of neighbor genes through cis-regulatory mechanisms (c). RNA-seq technologies allow to analyze individual HERV loci expression that may influence cellular gene expression (d). For example, HERVs may provide alternative promoters to neighbor genes, resulting in HERV-gene chimeric transcripts (1). HERV non-coding RNAs (ncRNAs) can also be detected through RNA-seq approaches (2).

ChIP-Seq can be also used to study the interaction of DNA with other proteins further than those composing chromatin, and to predict, for example, the interaction between LTRs and Transcription Factors (TFs). Indeed, due to the presence of enhancers into their LTR sequences, HERVs and solo LTRs integrated near to a cellular gene can enhance the expression of the gene through cis-regulatory mechanisms (Figure 5c) [28].

The HERV integrations may also interfere with the normal gene functions by providing alternative promoters, polyadenylation signals, and sites for splicing [28,125,126]. These mechanisms of interference can result in HERV-gene chimeric transcripts, which may be computationally reconstructed by applying RNA-seq to the transcriptome, and then using software that detects reads mapping in splice junctions [125]. HERV non-coding RNAs (ncRNAs) can also be detected through RNA-seq approaches, and the results obtained could be used as a start for studies of HERV mediated cis-regulation of cellular genes (Figure 5d) [28,93].

1.3.3 Identification of expressed and modulated HERV loci

The study of HERV implication in human physiopathology is also based on the analysis of the expression of each HERV locus, their modulation in different tissues, and pathological diseases (Figure 6) [5,113]. Indeed, the causal relationship between HERVs and diseases is still not clear, and most studies have not been able to identify the expression of individual loci, but of entire families [43,127]. The application of RNA-seq to the HERV transcriptome can provide essential information about the transcript contribution of the different families, the expression levels of the loci, and their localization in the human genome. For example, RNA-seq pipeline for differential expression can show over-expression or under-expression of HERVs in tissues [76], diseases [87], and medical treatments [84]. The data obtained with this approach can be used as a start for understanding the importance of HERV transcripts, their possible translation into proteins, but also for finding HERV biomarkers [44]. Finally, bioinformatics pipelines for metagenomics allow to study HERV contribution to human intestinal virome [128].

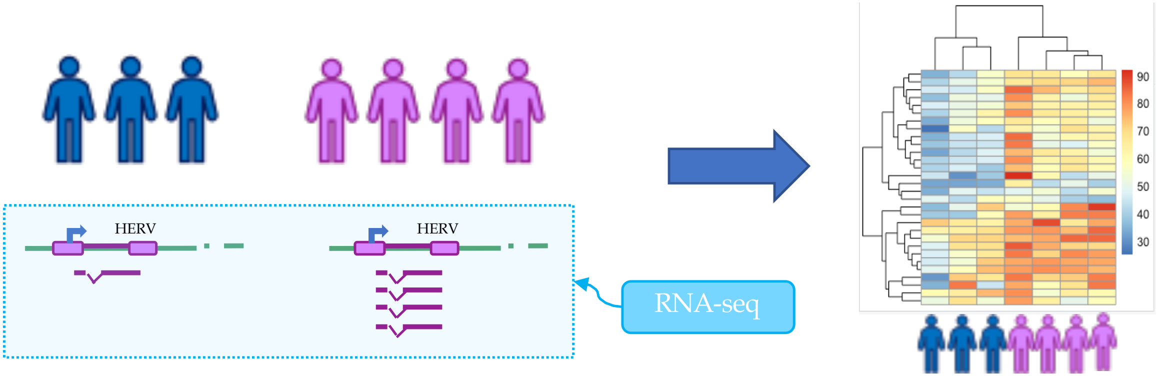


Figure 6. Example of possible application of High-Throughput sequencing for the identification of expressed and modulated HERV loci. The application of RNA-seq to the HERV transcriptome can provide the expression levels of the individual loci. Differential expression analyses can show HERVs modulation in different conditions, for example healthy controls versus diseases**.**

2. Aim of this study

Bioinformatics, together with the application of novel high-throughput sequencing technologies, have allowed great improvements in the study of HERVs. Indeed, these approaches can be used to obtain insights about the individual HERV loci from different point of views. For example, it is possible to collect the coordinates of HERV elements in the human genome, then to analyze and characterize them, and finally to use these data to study the expression and modulation of each locus (Figure 7).

This work takes advantage on two previous works of identification and classification of HERV elements [7,10], and it has the dual aim i) to characterize one HERV group, HML-6, and ii) to analyze pattern of HERV expression in specific contexts.

As a first step, we wanted to provide a comprehensive characterization of the HML-6 group members, which can be useful to direct further and more detailed studies of expression. We performed phylogenetic and structural analyses of the HML-6 internal sequences and LTRs, to asses similarity and heterogeneity between the HML-6 elements. Moreover, to provide updated information on the HML-6 presence in the human genome, we checked for their genomic context of insertion, which is essential to understand their potential role in physiological and pathological.

As a second step, we focused our attention on the HERV expression in PBMCs, and we analyzed their modulation in an immunity context. We studied the HERV expression patterns in two different *in vivo* models, using data from: i) participants to a clinical trial of healthy persons injected with LPS and ii) individuals being administered inactivated hantaan virus vaccine (Hantavax™). During the first analysis, the injection of high dose of LPS strongly mimicked the activation of the innate immune response mediated by bacteria. In this context we tried to understand the HERV activation over the first steps of the innate immunity. Instead, the vaccine administrations mimicked the activation of immunity triggered by the virus in the vaccine. The transcriptome data available for this vaccine study include PBMCs from individuals after the second, third and fourth administrations, giving hence us the opportunity to analyze the HERV modulation over the various steps of adaptive immunity.

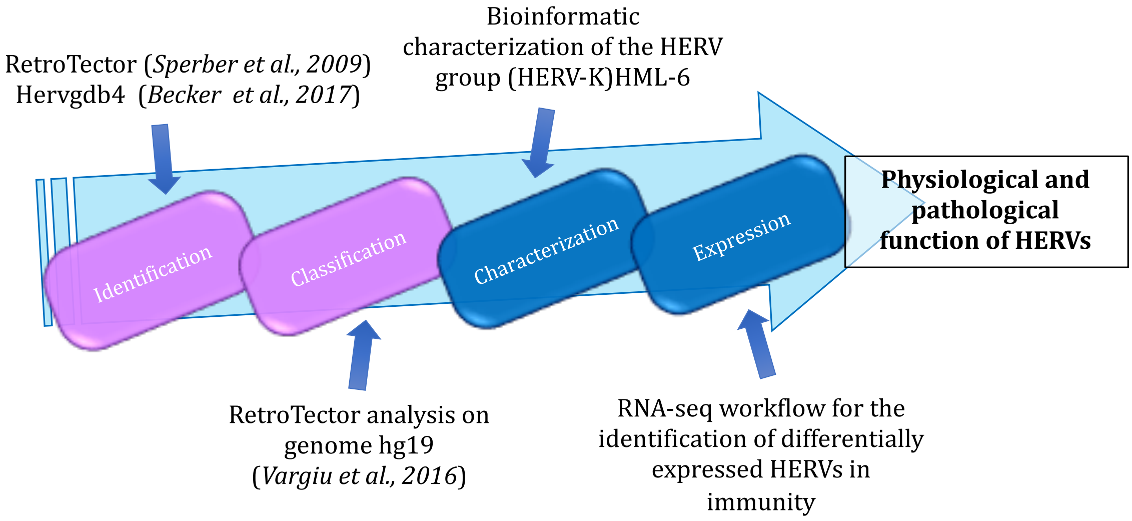

RetroTector (*Sperber et al., 2009*)
Hervgdb4 (*Becker et al., 2017*)
Bioinformatic characterization of the HERV group (HERV-K)HML-6
Identification
Classification
Characterization
Expression
Physiological and pathological function of HERVs
RetroTector analysis on genome hg19 (*Vargiu et al., 2016*)
RNA-seq workflow for the identification of differentially expressed HERVs in immunity

Figure 7. Schematic representation of workflow of this thesis work.

3. Material and methods

3.1 HML-6 group characterization

The HML-6 sequences were collected from the human genome assembly GRCh37/hg19 both by employing the RetroTector analysis on GRCh37/hg19 assembly and by retrieving chromosome coordinates in the UCSC Genome Browser database [129,130], using assembled LTR3A-HERVK3-LTR3A consensus sequences from Dfam database as BLAT query [131]. Elements obtained from both strategies were combined and the identity of the HML-6 sequences was confirmed by multiple alignments with respect to the assembled HERVK3 consensus sequence. We estimated the expected distribution of HML-6 loci in each chromosome by using the formula: $e=Cl*66/Tl$, where e is the number of expected integration in the chromosome, Cl is the chromosome length, 66 is the total amount of HML-6 loci in human genome hg19 and Tl is the sum of all chromosome lengths.

Using the LTR3, LTR3A, LTRB and LTR3B_v consensus sequences from Dfam as queries for a BLAT search, we collected the HML-6 solitary LTRs. The coordinates have been compared in order to exclude replicates. A consensus nucleotide alignment of the internal sequences has been created with MCoffee form the TCoffee package version 12.00.7fb08c2 [54]. The integrity of each HML-6 element was analyzed as compared to the assembled HERVK3 consensus sequence from Dfam database [131]. The genomic structure was furthermore defined by using the RetroTector algorithm [11] in ReTe online. Additional multiple alignments were performed with MAFFT online, version 7 [132], for the inspection of LTRs composition with respect to the LTR3, LTR3A, LTRB and LTR3B_v consensus sequences from Dfam. The obtained alignments were visualized on the Geneious bioinformatics software, version 8.1.3 [133]. The consensus sequences for different type 1a, type 1b and type 2 were generated from multiple alignments following the majority-based rule using the Geneious bioinformatics software, version 8.1.3.

We selected the Kimura model (K80) as the more appropriate for analyze the HML-6 internal sequence evolution with JmodelTest, version 2.1.10 [134]. Neighbor-joining phylogenetic trees were built with Mega Software [135], version 6.06, using pairwise deletion and p-distance method with 1000 bootstrap replications. Maximum likelihood trees were built with PhyML 3.0 online (http://www.atgc-montpellier.fr/phyml/), selecting the K80 model and 100 bootstrap replication [136]. The Gag amino acid sequences of the other HML consensus and of the exogenous retroviruses MPMV (P07567), MMTV and JSRV(P31622) were included in the analysis as control, as well as ZAM(O46144), used as an out-group. Phylogenetic trees of Rec sequences were built with Mega Software [135], version 6.06, using pairwise deletion and Poisson method. The HML-10 [17], HML-2 Rec (P61573, P61572, P61573, P61575, P61576, P61571, P61578), HTLV-1 Rex (Q85601) and HIV-1 Rev (P69718) amino acid sequences were included in the analysis.

Considering the HML-6 coevolution with the host genome and assuming a human genome substitution rate of 0.2% per nucleotide per million years, the time of integration of the HML-6 sequences (T) was estimated with the formula T = D/0.2, calculating the percentage of divergent nucleotides (D) between 150–350 nucleotide-length portions of *gag*, *pol* and *env* genes and a generated consensus for each type and subtype. The consensus sequences used in these analyses were generated with Geneious software from visually inspected multiple alignments and following the majority rule. The time of integration based on 5'- versus 3'- LTRs divergence was also evaluated, considering that each LTR of the same sequence accumulates mutations independently in the formula T=D/0.2/2 Divergence values were estimated on MEGA 6.06 [135], using pairwise deletion and Kimura 2-parameter model, and excluding CpG dinucleotides from the alignments. The final age of sequences was expressed, when possible, as the average value obtained from all methods, excluding those with a standard deviation >25%.

The PBS nucleotide sequences were analyzed and characterized through MAFFT multiple alignments, in comparison to the PBS reference sequences kindly provided by Professor Blomberg. The nucleocapsid Zinc-fingers, the Pro dUTPases, and the Pol G-patch amino-acid motifs were aligned using the MUSCLE algorithm in MEGA [135]. All the analyses were visualized on the Geneious platform. The composition of PBS and structural features was represented with a WebLogo (http://weblogo.berkeley.edu).

The genomic context of the HML-6 elements was retrieved by analyzing their genomic coordinates on the Data Integrator tool in UCSC Genome Browser [129,130], selecting the Genes and Genes prediction track. Moreover, all the sequences were visualized on Genome Browser concurrently with the activation of GENECODE v24, RefSeq genes, ENCODE and Gtex annotations. The distances between HERV proviruses/solitary LTRs and human genes have been computed by using the function “distance” from the package GenomicRanges version 1.30.3 on RStudio (R version 3.4.4). Human genes coordinates were collected from GENCODE v24.

The conserved domains present in the sequences were identified by using the NCBI Conserved Domain Search software [137].

3.2 HERV and MaLR expression and modulation in PBMCs

We used RNA-seq datasets public available (GEO:GSE87290 and GEO:GSE120115). GSE87290 includes the transcriptome of PBMCs from healthy humans (n=15) before and after 1ng/kg LPS exposure. Specifically, whole blood RNA samples were collected at baseline and 2 hours post LPS stimulation. PBMCs were isolated from whole blood immediately after collection [58]. GSE120115 includes the transcriptome of PBMCs from healthy humans (n=19) 1 day before and 2 days after the 2nd, the 3rd and the 4th administration of Hantavax™ inactivated vaccine against hantaan virus.

We checked the quality of the RNA sequences by using FastQC Galaxy Version 0.72 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Low quality reads were trimmed with Trim Galore! V.0.4.3.1 (https://github.com/FelixKrueger/Trim-Galore).

All the mentioned analyses were done on Galaxy release_17.09 locally installed (http://galaxyproject.org/).

HISAT2 Galaxy Version 2.1.0 was used with default parameters to map reads to genome assembly hg38. We assessed the quality of the alignments by using the stats function of bamtools Galaxy Version 2.0.1 [138]. We counted the reads mapping to each *hervgdb4* fragment and human gene by using the "union" mode in htseq-count Galaxy Version 0.6.1galaxy3 [139], and *hervgdb4* database [10] and gencode.v27 [140] for respectively HERVs/MaLRs and gene annotations. All the mentioned analyses were done on Galaxy release_17.09 locally installed (http://galaxyproject.org/). We calculated the expression values of the expressed *hervgdb4* fragments and genes as Transcripts Per Million (TPM).

We selected all the genes and *hervgdb4* fragments with at least 1 count in at least 10 samples (LPS stimulation) or with at least 1 count in at least 26 samples (vaccine administration), conventionally considering them as expressed. We used the GenomicRanges v.1.30.3 [141] R (version 3.4.4) package to obtain the coordinates of the most intact HERV proviruses [7] from those of the expressed *hervgdb4* fragments.

The DESeq2.v.1.18.1 R (version 3.4.4) package [142] has been used to perform rlog normalization on human genes and HERV/MaLR raw counts. From the output of the normalization we extracted the HERV/MaLR rlog counts and we assessed the interpersonal variability through PCA and Heatmap. The PCA was built with the function plotPCA in DESeq2.v.1.18.1 and visualized by using ggplot 3.0.0 in R (version 3.4.4). The Heatmaps of the *hervgdb4* fragments with the higher average/standard deviation rlog counts across samples were built through the pheatmap 1.0.10 R (version 3.4.4) package, considering the correlation distance across samples in column. The dist function in DESeq2.v.1.18.1 was applied to the transpose of the rlog transformed count matrix to get Euclidean sample-to-sample distances; the heatmaps were built through the pheatmap 1.0.10 R (version 3.4.4) package.

We performed a differential expression analysis of the genes and *hervgdb4* fragments by using the DESeq2.v.1.18.1 R (version 3.4.4) package [142] and the filtered raw counts as input. During the analysis, false discovery rate/adjusted p-value were used for multiple test comparison according to the Benjamini-Hochberg procedure [143]. We used a threshold (FDR <= 0.01 and absolute values of $\log_2$ Fold Change >= 1) to identify the modulated elements. The differentially expressed *hervgdb4* fragments were visualized in a volcano plot built by using ggplot 3.0.0 in R (version 3.4.4). We sorted the *hervgdb4* fragments by adjusted p-values to recognize the mostly differentially expressed. We compared the TPM expression values with the filtered adjusted p-values (<= 0.01) to see the relative distributions. We manually checked for the presence of neighbor genes on ENSEMBL [144], within a 10-kb window of distance from the nearest-neighbor gene. For the 10 most modulated elements after LPS stimulations co-localized with human genes, we reconstructed the transcripts. The transcripts have been reconstructed by using Trinity-v2.5.1 [145] on a subset of reads mapping in genome ranges that included the HERVs and MaLRs, the human gene and a flanking region of 500 bp. The transcripts have been mapped to the genome assembly hg38 by gmap-2019.09.12 [146] and finally visualized in Integrative Genomics Viewer (IGV) [147]. The bound of the HERVs and MaLRs analyzed have been manually inspected through the sashimi-plot function of IGV.

4. Comprehensive characterization of the HERV-K(HML-6) group: overview of their structure, phylogeny and contribution to the human genome

4.1 Collection of 66 HML-6 loci in the human genome sequence

We collected the HML-6 sequences provided in *Vargiu et al.* 2016, in which the RetroTector (ReTe) analysis of the genome assembly hg19 allowed to identify and classify all the most intact HERV proviruses in our genome [7]. Moreover, we compared these coordinates and sequences with those obtained with Genome Browser BLAT search in genome assembly hg19, using as a query the LTR3A-HERVK3-LTR3A consensus sequence assembled from Dfam dataset. In particular, 2 HML-2 sequences - in locus 10q11.21 and 10q25.1 - were detected only by ReTe and provided in Vargiu *et al.*, but we were not able to find them by BLAT search. Similarly, 2 sequences in locus 5q13.2, showing 100% identity and flanked by identical region, were both detected by BLAT, even if only one sequence in locus 5q13.2 was previously reported in Vargiu. Therefore, through this integrated search approach, we obtained the genomic coordinates of 66 HML-6 sequences (Table 2). We named the HML-6 elements in conformity with their genomic localization, and in the case of presence of multiple sequences within the same genomic locus, we unequivocally indicated the sequence order with alphabetical letters.

When we analyzed the distribution of the HML-6 insertion, almost all chromosomes showed an apparent random distribution of HML-6 loci, in the sense that the number of sequences was approximately proportional to the chromosome size. The exceptions were chromosome 19 a chromosome Y, in which we detected more HML-6 proviruses than expected. An overall chi-square test including all chromosomes indicated a non-random distribution (p<0.0001) of HML-6 loci, with a very prominent

Table 2. HML-6 proviral sequences and their localization in the human genome GRCh37/hg19 assembly**.**

Locus	Chr	Strand	Start	End	LTR type	Length	Subtype
1p21.1	1	(+)	103298830	103306681	LTR3	7851	1a
1q25.2	1	(-)	179406261	179412404	LTR3A	6143	1a
2q14.23	2	(-)	128372842	128376247	LTR3B_v	3405	2
2q22.1	2	(-)	136829388	136834831	LTR3A	5443	1a
3p25.1	3	(+)	14266558	14271762	LTR3A	5204	1a
3p21.31a	3	(-)	46087646	46095966	LTR3A	8320	1a
3p21.31b	3	(+)	46468034	46475121	LTR3A	7087	1a
4p14	4	(-)	39540876	39545998	LTR3B	5122	1b
4q13.2	4	(-)	69610304	69616956	LTR3B	6652	1b
4q13.3	4	(+)	71418184	71420406	LTR3B_v	2222	2
4q21.1	4	(-)	78313436	78321358	LTR3	7922	1a
5p14.1	5	(+)	24649773	24654829	LTR3A	5056	1a
5q13.2a	5	(+)	69641005	69643229	only 3'LTR3B_v	2224	2
5q13.2b	5	(+)	69958435	69960659	only 3'LTR3B_v	2224	2
5q13.2c	5	(+)	70867724	70874228	LTR3A	6504	1a
6p22.2	6	(-)	26288250	26296494	LTR3B	8244	1b
6p21.32a	6	(-)	32443272	32447375	only 5'LTR3	4103	1a
6p21.32b	6	(-)	32527497	32535122	only 5'LTR3	7625	1a
7q36.1	7	(-)	150279386	150283313	LTR3B_v	3927	2
8q11.1	8	(+)	47395171	47403000	LTR3A	7829	1a
10q11.21	10	(-)	43796372	43788582	LTR3B_v	7790	2
10q11.21	10	(+)	45774424	45782353	LTR3	7929	1a
10q25.1	10	(+)	110488980	110492726	LTR3B_v	3746	2
11p15.4	11	(-)	7920872	7927779	LTR3A	6907	1a
11q12.3a	11	(-)	61817251	61823827	LTR3A	6576	1a
11q12.3b	11	(+)	62019229	62021808	only 5'LTR3	2579	1a
11q23.2	11	(-)	112795351	112800806	LTR3A	5455	1a
12q24.12	12	(-)	112253979	112263658	LTR3A	9679	1a
14q12	14	(+)	28879914	28890437	only 3'LTR3A	10523	1a
14q24.2	14	(+)	70278180	70282740	LTR3A	4560	1a
16p11.2	16	(-)	30627018	30635602	LTR3B	8584	1b
16p11.1	16	(+)	34750975	34758850	LTR3B	7875	1b
17q21.31	17	(+)	41949365	41952180	LTR3B_v	2815	2
17q25.1	17	(+)	72580505	72583070	only 5'LTR3B	2565	1b
19p13.2a	19	(-)	9618707	9625736	LTR3	7029	1a
19p13.2b	19	(-)	11964097	11971799	LTR3	7702	1a
19p12a	19	(+)	21416592	21420867	LTR3B_v	4275	2
19p12b	19	(-)	21968952	21975023	only 3'LTR3	6071	1a

19p12c	19	(+)	22376043	22379350	only 3'LTR3	3307	1a
19p12d	19	(+)	24047631	24054030	LTR3	6399	1a
19q13.41a	19	(-)	52307949	52315192	LTR3A	7243	1a
19q13.41b	19	(+)	52479404	52484683	LTR3A	5279	1a
19q13.41c	19	(-)	52913436	52917986	---	4550	1a
19q13.41d	19	(-)	52978909	52981951	LTR3A	3042	1a
19q13.41e	19	(-)	53487788	53492829	only 3'LTR3B	5041	1b
19q13.43a	19	(+)	58023984	58029856	LTR3B	5872	1b
19q13.43b	19	(+)	58817037	58826633	LTR3B	9596	1b
20p13	20	(-)	1377446	1383348	LTR3A	5902	1a
20p11.21	20	(+)	25374769	25383907	LTR3A	9138	1a
Xp11.22	X	(+)	53188296	53193008	LTR3	4712	1a
Xp11.21	X	(-)	57129414	57135829	LTR3A	6415	1a
Xq13.2	X	(+)	73397834	73402327	only 3'LTR3A	4493	1a
Xq27.1	X	(-)	140290665	140293656	only 5'LTR3B	2991	1b
Yq11.221	Y	(+)	19443452	19448989	LTR3A	5537	1a
Yq11.222a	Y	(+)	19958329	19963018	only 3'LTR3B	4689	1b
Yq11.222b	Y	(+)	20051008	20055589	only 3'LTR3B	4581	1b
Yq11.222c	Y	(-)	20074149	20078731	only 3'LTR3B	4582	1b
Yq11.222d	Y	(-)	20216759	20221448	only 3'LTR3B	4689	1b
Yq11.223a	Y	(-)	25964947	25969633	only 3'LTR3B	4686	1b
Yq11.223b	Y	(+)	26162248	26166935	only 3'LTR3B	4687	1b
Yq11.23a	Y	(-)	26263056	26267731	only 3'LTR3B	4675	1b
Yq11.23b	Y	(-)	26277752	26282358	only 3'LTR3B	4606	1b
Yq11.23c	Y	(+)	27680077	27684680	only 3'LTR3B	4603	1b
Yq11.23d	Y	(-)	27694700	27699373	only 3'LTR3B	4673	1b
Yq11.23e	Y	(+)	27795496	27800183	only 3'LTR3B	4687	1b
Yq11.23f	Y	(+)	27992754	27997440	only 3'LTR3B	4686	1b

contribution of chi-squares calculated for chromosomes 19 and Y (Figure 8a). To confirm this finding, we also performed all comparisons between each pair of chromosomes and, owing to the large number of tests involved (n=276), filtered p values by the Benjamini-Hochberg procedure [143] to maintain a cumulative probability of false discoveries in all tests lower than 5%. Data confirmed a significant enrichment on chromosomes 19 and Y (Figure 8b).

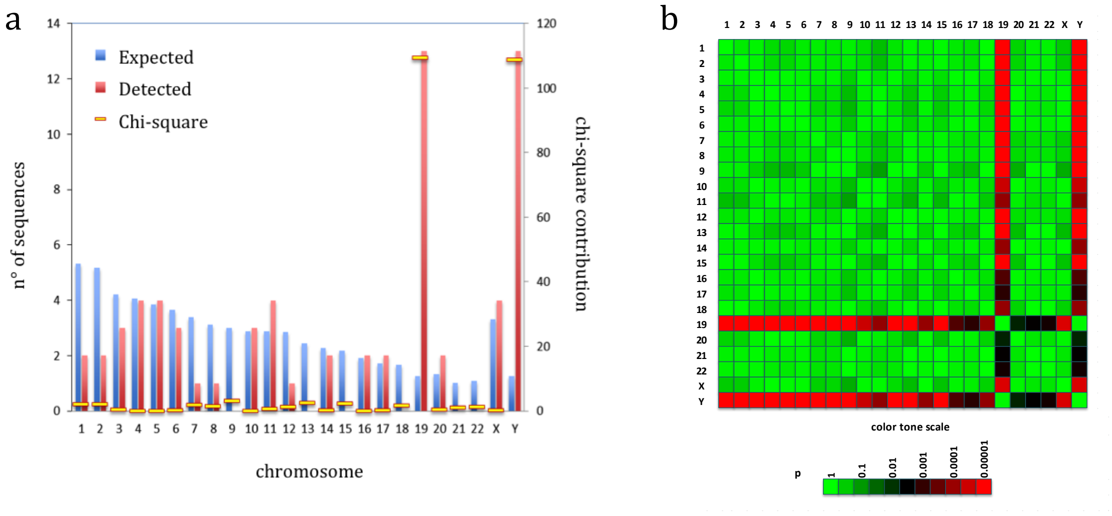


Figure 8. Histogram of the chromosomal expected and detected distribution of HML-6 proviruses. The enrichment in chromosomes 19 and Y is particularly clear, as also highlighted by their contribution to the chi-square distribution (a). The statistic of this enrichment is shown in the heatmap of pvalues obtained from the chi-square for the difference in the frequency of HML-6 proviruses insertions calculated between each pair of chromosomes (b). All pvalues obtained from the comparison of chromosome 19 and Y with all other chromosomes were statistically significant, with a cumulative probability of false discoveries less than 5%, as assessed by the Benjamini-Hockberg procedure. Conversely, all other comparisons were not statistically significant.

4.2 Phylogenetic analyses and subtype classification of HML-6 proviral internal sequences

In order to characterize the structure of each single provirus, we created a consensus multiple sequence alignment of i) all the 66 HML-6 internal sequences and ii) the consensus sequence HERVK3 from Dfam. Next, we performed a Neighbor-Joining (NJ) analysis of the created consensus alignment with the Kimura model test. The resulting tree revealed the presence of two main clusters that we named type 1 and 2, including 55 and 11 elements, respectively (Figure 9a). Moreover, type 1 elements showed an additional internal subdivision in two further clusters of 35 and 20 elements that we named type 1a and type 1b, respectively. To better understand the meaning of this phylogenetic information, we implemented the analysis by creating a maximum likelihood tree, selecting the K80 model for the phylogenetic analysis (Figure 9b). Also in this case we obtained a similar result, as we were able to identify the two main clusters with type 1 and 2 HML-6 proviruses, and the type 1 subdivision in type 1a and 1b. Anyway, loci Xq27.1 and 17q25.1 were assigned to type 2 cluster by the NJ analysis while to the type 1b cluster by the ML analysis. Subsequently, phylogeny and evolutionary relations of the HML-6 group was investigated with respect to the others HML elements. We generated a majority-rule consensus sequence for the *gag* gene of each subgroup, selecting this gene due to the fact that it was the most conserved within the HML-6 group. Then, we performed a NJ analysis of amino acid sequences of Gag, by using our translated consensus and the Gag consensus sequences of the others HML groups [7], whereas the Gag sequences of the exogenous Beta-retroviruses MMTV, MPMV and JSRV, and the Gag sequence of ZAM *drosophila* endogenous retrovirus were used as out-groups. This analysis confirmed that all analyzed sequences belonged indeed to the HML-6 group (Figure 10). Interestingly, all the HML-6 Gag consensus sequences grouped together outside the HML clade, suggesting that HML-6 may represent an intermediate group between the HERV-K and the other MMTV-related clades, in agreement to what has been reported in other works [100,148,149].

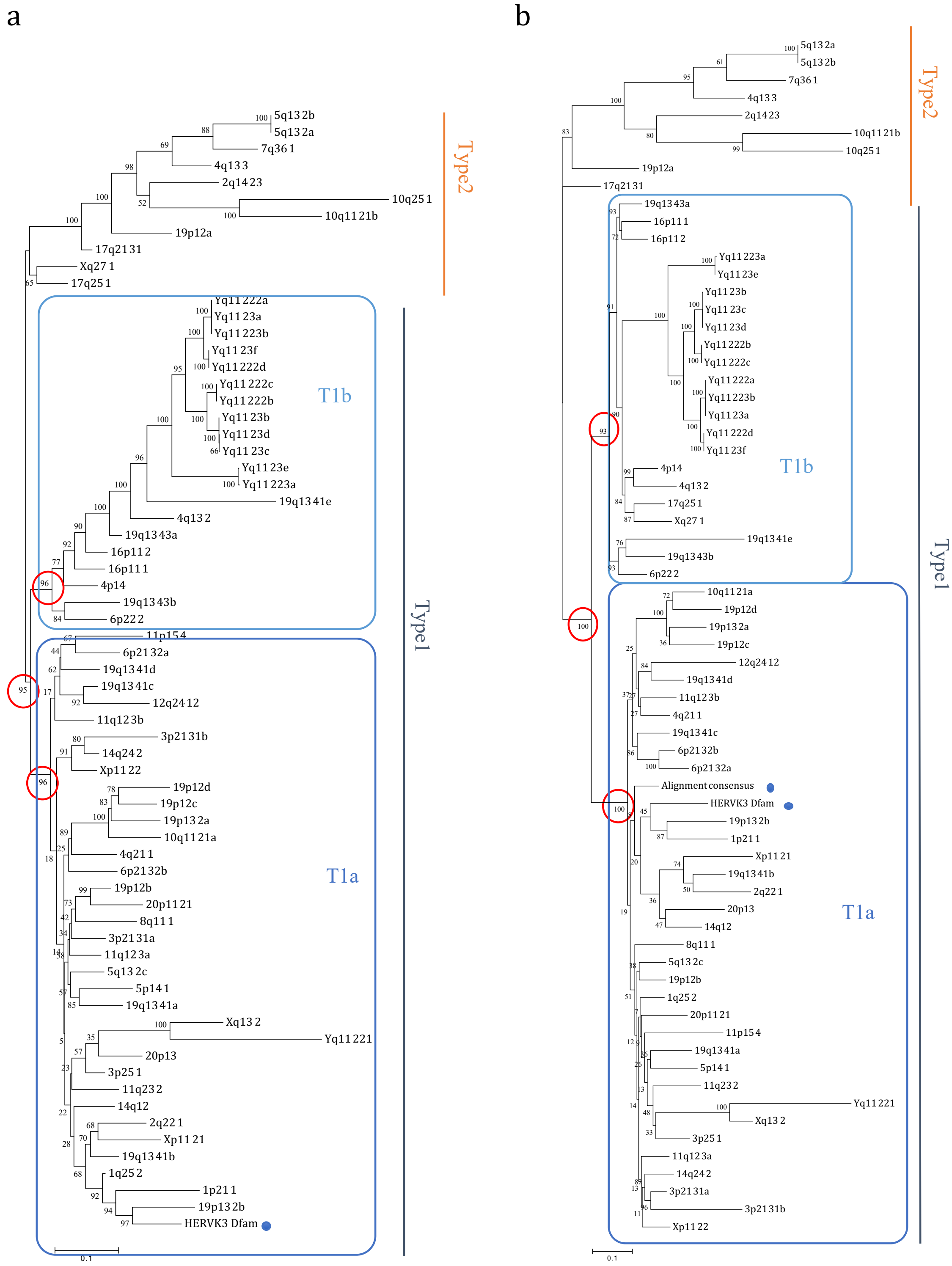


Figure 9. Phylogenetic analysis of the internal sequences. The phylogeny of was investigated by using Neighbor Joining method and the Kimura-2-parameter model (a) and by performing a ML analysis and K80 model test (b) of a consensus alignment of internal sequences. The two intragroup clusters (Type 1 and 2) are indicated with blue and red lines, respectively, whereas the additional distribution of type 1 elements in two subtypes (1a and 1b) is indicated by squares.

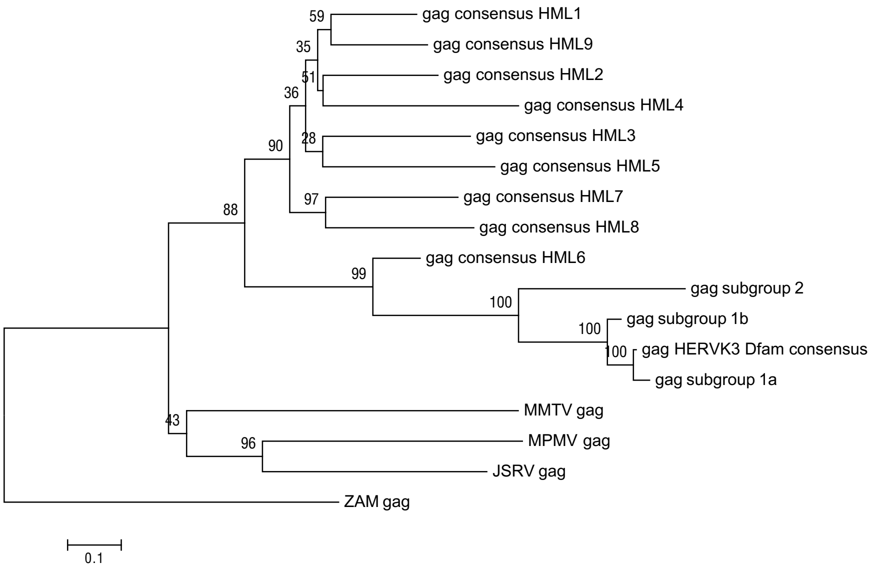


Figure 10. Phylogenetic tree of the HML gag consensus sequences. The phylogeny of HML-6 group was investigated by performing a NJ analysis of amino acid sequences of gag. Our consensus sequences for the gag gene of each subgroup and the gag consensus sequences of the others HML groups are included in this tree. The gag sequences of the exogenous betaretroviruses MMTV, MPMV and JSRV, and gag sequence of ZAM, endogenous retrovirus in *drosophila* were used as out-groups. The phylogenetic tree has been built by using the Bootstrap method with 1000 replicates.

4.3 Structural characterization of HML-6 proviral sequences

The Dfam assembled HML-6 consensus sequence shows a typical proviral genome structure, in which 5'- and 3'-LTRs flank the *gag*, *pro*, *pol* and *env* genes, encoding the structural proteins and the essential enzymes [99]. The *gag* gene (position 205–1850) encodes MA, CA and NC elements; the *pro* gene (1666–2621) encodes the Pro enzyme; the *pol* gene (2578-5276) determines the production of RT and IN; and the *env* gene (5225-7166) encodes SU and TM proteins. In addition, the analysis of conserved domains allowed identifying a predicted Rec domain between *pol* and *env* (5272-5445).

We hence attempted to define the structural characteristics of the HML-6 proviral types, annotating all the insertions and deletions within the internal sequences with respect to the consensus (Figure 11). In general, as compared to the consensus sequence, which was 7166 bp in length excluding LTRs, the overall average length of 4918 bp was below the expectation. In fact, only 9 elements maintained a complete structure, whereas the majority of sequences were incomplete due to the lack of large viral portions (Figure 11).

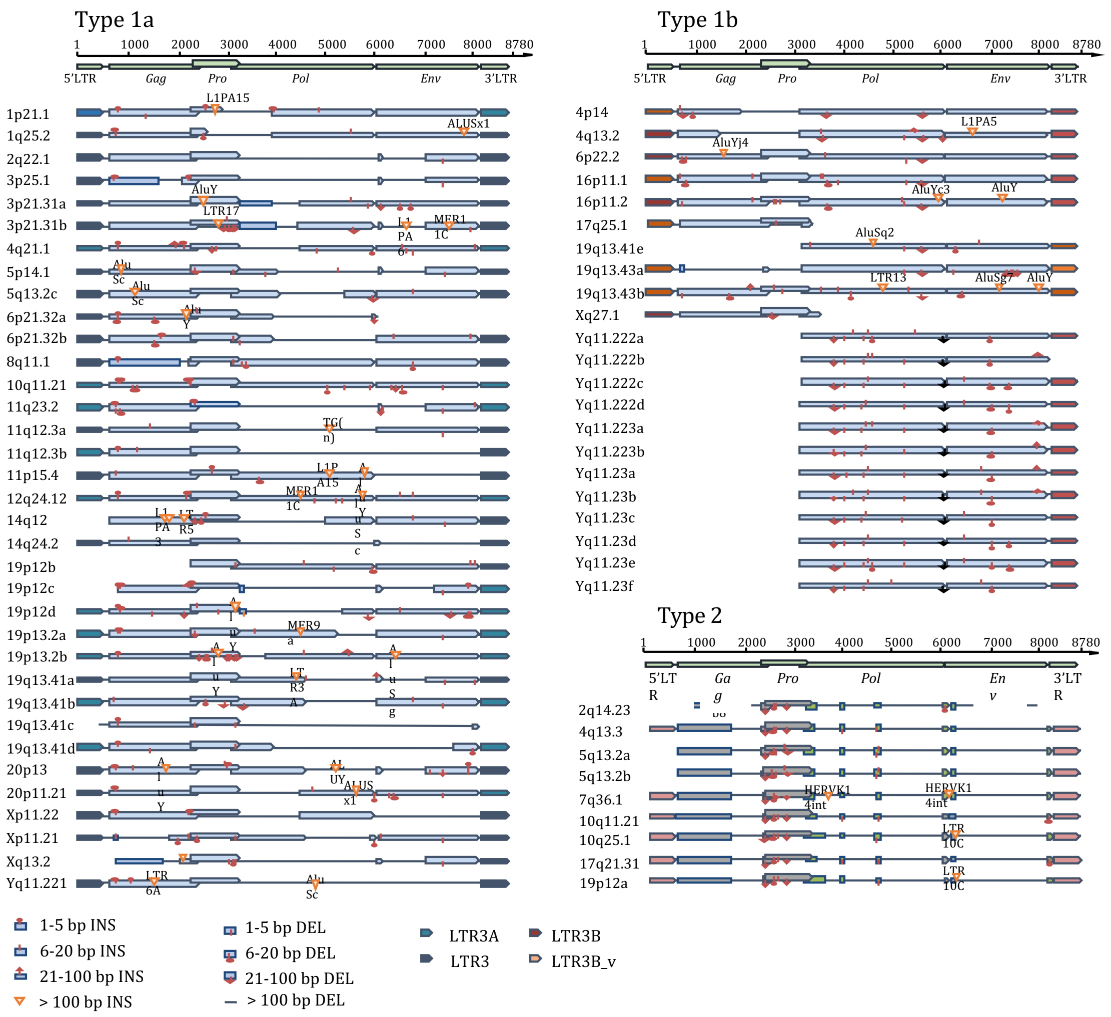


Figure 11. Structural characterization of 66 HML-6 sequences. Nucleotides insertions and deletions of each HML-6 nucleotide sequence has been annotated by comparison with the HML-6 consensus sequence from DFAM (a). Type 2 elements (in green) showed common nucleotide deletion partially or totally corresponding to *pro* (nt 2736-3120) *pol* (nt 3625-3972, 4133-4561, 4653-6201) and *env* (nt 6596-8488) protein domains, as shown in the alignments of type 1 and 2 consensus sequences (b).

We hence annotated these variations, observing that a number of them are shared between elements of the same subgroup: i) 51% of subgroup 1a elements lacked nucleotides 2727–4060 within the *pol* gene (RT portion), whereas the 57% lacked nucleotides 5176–6185 within the *env* gene (Rec and SU portion); ii) the *gag* and *pro* genes were completely missing in 69% of subgroup 1b elements, and, moreover, a deletion of 5236–5333 (Rec portion) nucleotides within the *env* gene was also found in the 69% of these sequences; iii) the viral portions between nucleotides 717–1000, 2239-2919, 3078-3492, 3592-5071 and 5249-7120 were deleted in all the subgroup 2 elements, resulting in the partial deletion of the *gag* (p17 portion) and *pro* genes, and in the complete deletion of the *pol* and *env* genes. We consequently summarized a consensus structure for each subgroup (Figure 12).

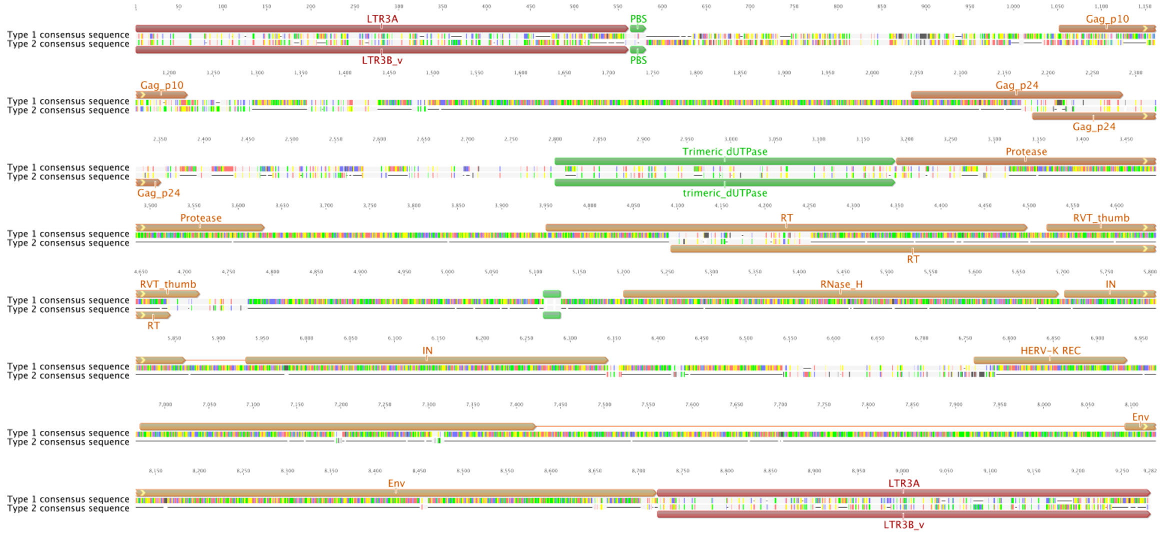


Figure 12. HML-6 type 1 and type 2 consensus sequences. Type 2 elements showed common nucleotide deletion partially or totally corresponding to *pro, pol* and *env* protein domains, as shown in the alignments of type 1 and 2 consensus sequences.

In addition, we annotated all minor insertions and deletions, in order to define not only the subgroup overall identity but also the singularity of each HML-6 sequence (Figure 11). Such a detailed structural characterization may hence provide a specific background for the structural investigation of single HML-6 loci and the unequivocal match to their eventual expression products.

4.4 Phylogenetic analysis of individual HML6 retroviral genes

To further verify the previous phylogenetic and structural studies, we then performed NJ analyses for the individual HML-6 *gag*, *pro*, *pol* and *env* genes that confirmed the presence of two main proviral types (1 and 2), as well as the additional subdivision within the type 1 (1a and 1b) for all genes (Figure 13). Moreover, we inspected the characteristics of the newly identified HML-6 Rec putative domain. Firstly, we used the ERVK3-1 *rec* nucleotide sequence as a reference for a multiple alignment, finding that 39 of the 66 identified HML-6 elements included the *rec* sequence within their *env* gene. Then, we created a multiple alignment of the predicted Rec amino acid sequences, finding a full-length Rec putative domain within 23 HML-6 loci, while 16 loci showed an incomplete Rec domain due to the presence of several deletions (Figure 14a). We hence built a consensus sequence from the multiple alignment a NJ phylogenetic tree of 23 full-length HML-6 Rec amino acid sequences (Figure 14b). We included as reference sequences: i) 7 HML-2 Rec amino acid sequences reported in Uniprot; ii) the recently described HML-10 Rec consensus amino acid sequence [17]; iii) the amino acid sequence of the functional homologue HIV-1 Rev; iv) the amino acid sequence of the functional homologue HTLV-1 Rex (see materials and methods for the correspondent accession numbers). Remarkably, the Rec NJ tree showed a high phylogenetic relationship between HML-6 and HML-2 Rec putative proteins.

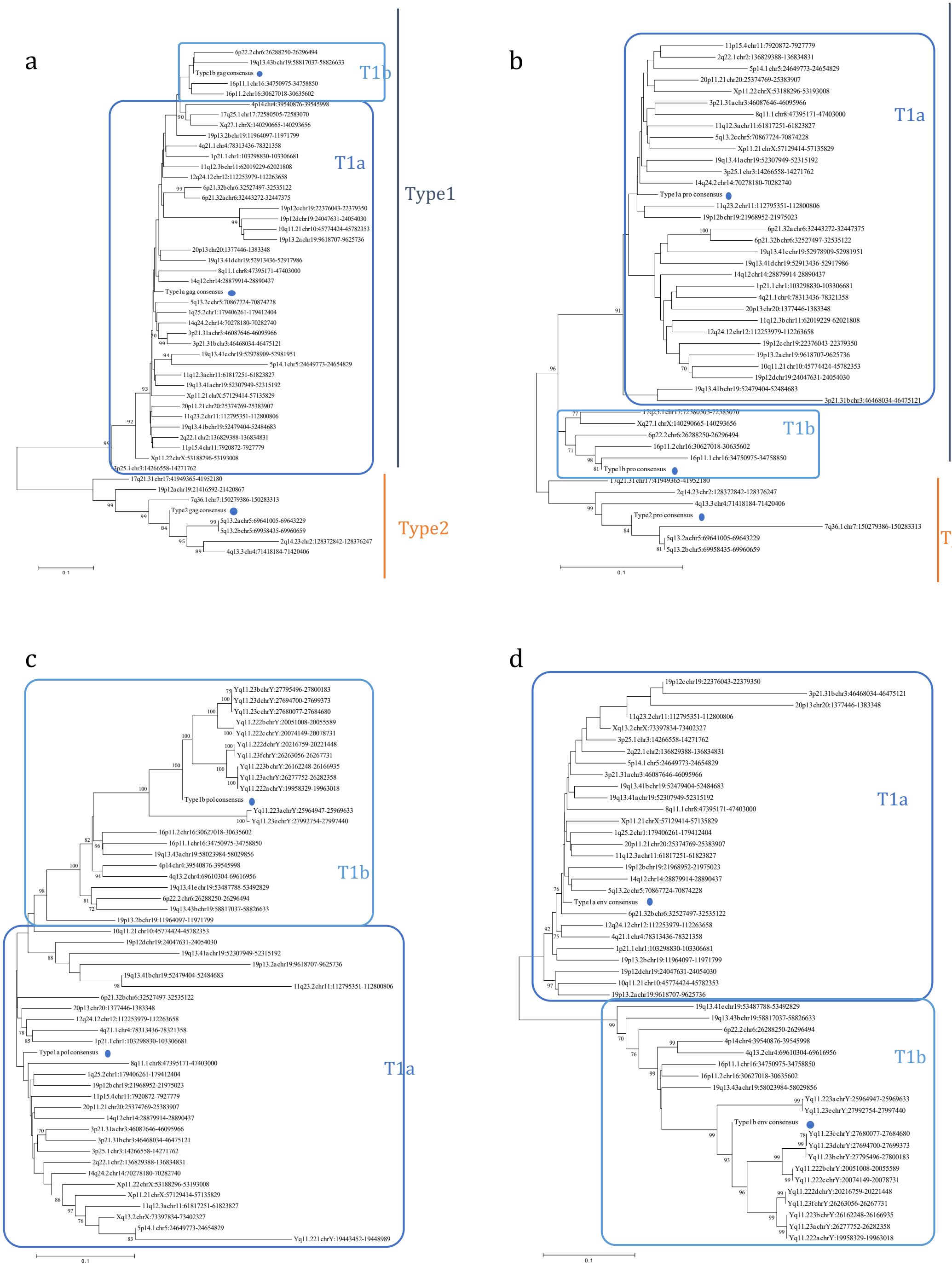


Figure 13. Phylogenetic analysis of the HML-6 nucleotide sequences of *gag* (a), *pro* (b), *pol* (c) and *env* (d) genes. When the genes are presents, two intragroup type (1 and 2), are indicates by blue and red lines. The three intragroup consensus sequences of the genes are also included in the analysis and indicated with a dot. The evolutionary relationship has been ascertained by using Neighbor Joining method and the Kimura-2-parameter model, the phylogenetic tree has been built by using the Bootstrap method with 1000 replicates.

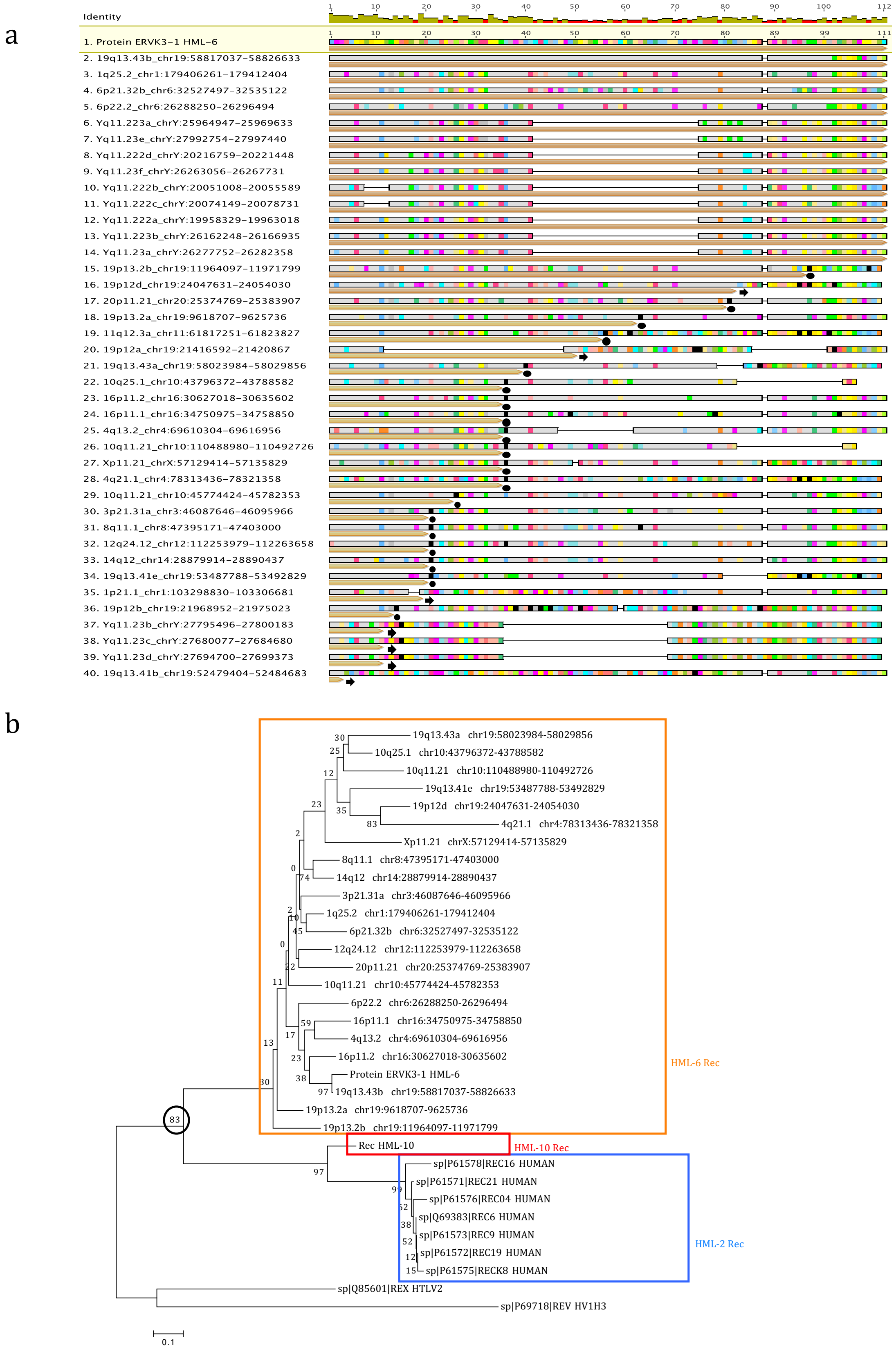

a
Identity
1. Protein ERVK3-1 HML-6
2. 19q13.43b_chr19:58817037-58826633
3. 1q25.2_chr1:179406261-179412404
4. 6p21.32b_chr6:32527497-32535122
5. 6p22.2_chr6:26288250-26296494
6. Yq11.223a_chrY:25964947-25969633
7. Yq11.23e_chrY:27992754-27997440
8. Yq11.222d_chrY:20216759-20221448
9. Yq11.23f_chrY:26263056-26267731
10. Yq11.222b_chrY:20051008-20055589
11. Yq11.222c_chrY:20074149-20078731
12. Yq11.222a_chrY:19958329-19963018
13. Yq11.223b_chrY:26162248-26166935
14. Yq11.23a_chrY:26277752-26282358
15. 19p13.2b_chr19:11964097-11971799
16. 19p12d_chr19:24047631-24054030
17. 20p11.21_chr20:25374769-25383907
18. 19p13.2a_chr19:9618707-9625736
19. 11q12.3a_chr11:61817251-61823827
20. 19p12a_chr19:21416592-21420867
21. 19q13.43a_chr19:58023984-58029856
22. 10q25.1_chr10:43796372-43788582
23. 16p11.2_chr16:30627018-30635602
24. 16p11.1_chr16:34750975-34758850
25. 4q13.2_chr4:69610304-69616956
26. 10q11.21_chr10:110488980-110492726
27. Xp11.21_chrX:57129414-57135829
28. 4q21.1_chr4:78313436-78321358
29. 10q11.21_chr10:45774424-45782353
30. 3p21.31a_chr3:46087646-46095966
31. 8q11.1_chr8:47395171-47403000
32. 12q24.12_chr12:112253979-112263658
33. 14q12_chr14:28879914-28890437
34. 19q13.41e_chr19:53487788-53492829
35. 1p21.1_chr1:103298830-103306681
36. 19p12b_chr19:21968952-21975023
37. Yq11.23b_chrY:27795496-27800183
38. Yq11.23c_chrY:27680077-27684680
39. Yq11.23d_chrY:27694700-27699373
40. 19q13.41b_chr19:52479404-52484683
1 10 20 30 40 50 60 70 80 90 100 112
1 10 20 30 40 50 60 70 80 89 99 111
b
19q13.43a chr19:58023984-58029856
10q25.1 chr10:43796372-43788582
10q11.21 chr10:110488980-110492726
19q13.41e chr19:53487788-53492829
19p12d chr19:24047631-24054030
4q21.1 chr4:78313436-78321358
Xp11.21 chrX:57129414-57135829
8q11.1 chr8:47395171-47403000
14q12 chr14:28879914-28890437
3p21.31a chr3:46087646-46095966
1q25.2 chr1:179406261-179412404
6p21.32b chr6:32527497-32535122
12q24.12 chr12:112253979-112263658
20p11.21 chr20:25374769-25383907
10q11.21 chr10:45774424-45782353
6p22.2 chr6:26288250-26296494
16p11.1 chr16:34750975-34758850
4q13.2 chr4:69610304-69616956
16p11.2 chr16:30627018-30635602
Protein ERVK3-1 HML-6
19q13.43b chr19:58817037-58826633
19p13.2a chr19:9618707-9625736
19p13.2b chr19:11964097-11971799
HML-6 Rec
Rec HML-10
HML-10 Rec
sp|P61578|REC16 HUMAN
sp|P61571|REC21 HUMAN
sp|P61576|REC04 HUMAN
sp|Q69383|REC6 HUMAN
sp|P61573|REC9 HUMAN
sp|P61572|REC19 HUMAN
sp|P61575|RECK8 HUMAN
HML-2 Rec
sp|Q85601|REX HTLV2
sp|P69718|REV HV1H3
83
97
0.1

Figure 14. Multiple alignment and phylogenetic relationships of HML-6 Rec domains. Multiple alignment of the HML-6 Rec amino acid sequences with the protein ERVK3-1 used as consensus (a). The colors into the sequences show disagreements in the alignments; the black lines represent the deletion. The ORFs are indicated with orange arrows, eventually stopping in correspondence of stop codons (black dots) or frame-shift mutation (black

arrows). The relationship between the 4 best preserved HML-6 Rec domains and the known HML-2 and HML-10 Rec domains is shown in a Phylogenetic tree (b). This relationship has been ascertained by using Neighbor Joining method and the Kimura-2-parameter model, 1000 bootstrap replicates.

Secondly, to investigate their possible relevance, we also analyzed the integrity of the Open Reading Frames (ORFs) as compared to the ERVK3-1 Rec amino acid sequence (Figure 14a), observing that 4 out of 23 ORFs have a predicted intact coding structure devoid of premature stop codons and frameshifts. Thirdly, we focused on the 4 HML-6 Rec putative amino acid sequences with predicted intact ORFs, searching the functional domains involved into the nuclear localization (NLS) and export (NES) as described for the HML-2 and HML-10 Rec domains [17,96]. Results showed that the HML-6 Rec putative domain harbors a conserved NES domain as reported for the HML-2 and HML-10 Rec, but not the NLS domains reported for the HML-2 Rec, as well as for HIV Rev and HTLV Rex (data not shown). In addition, we investigated the presence of a putative Rec Responsive Element (RcRe) searching within the *env* sequence or within the LTRs similarities with the reported HML-2 RcRe and the HIV RRE. This analysis did not allow us to find any putative HML-6 RcRe element, however we cannot exclude the presence of a RcRe element characterized by different structures.

4.5 Characterization and phylogenetic analysis of LTRs and time of insertion

The HML-6 group has been associated with four different types of LTRs, named LTR3, LTR3A, LTR3B and LTR3B_v according to the Repeat Masker annotations. We hence attempted to characterize the structure and phylogeny of LTR sequences, inspecting in particular if the different LTR types were associated with specific proviral types.

Firstly, we performed a nucleotide sequence comparison between the 4 LTR types observing that LTR3A appeared to be a 3'-end extension variant of LTR3 (46 nt longer), while LTR3B_v seemed to be a 3'-end extension variant of LTR3B (62 nt longer) (Figure 15).

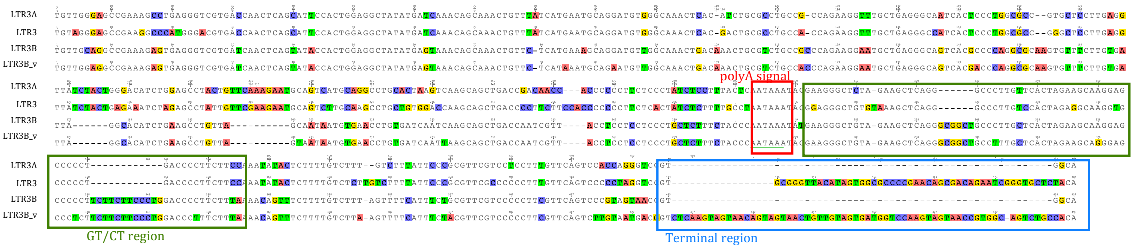


Figure 15. Structural characterization of HML-6 LTRs. The alignments between the LTR3A, LTR3, LTR3B and LTR3B_v Dfam consensus sequences are showed.

Secondly, to better identify the differences between LTRs, to have a comparison and to verify the Repeat Masker annotations, we built a NJ tree of all HML-6 proviral LTRs, which showed the presence of three clusters of sequences, one including LTR3A and LTR3, one including LTR3B and one including LTR3B_v elements (Figure 16). Hence, we asked whether there was any association between the type of LTR and the different HML-6 types. Interestingly, results showed that type 1a elements were associated with only LTR3 and LTR3A, type 1b sequences only occurred with LTR3B, and LTR3B_v were only related to type 2 members (Table 2, Figure 11).

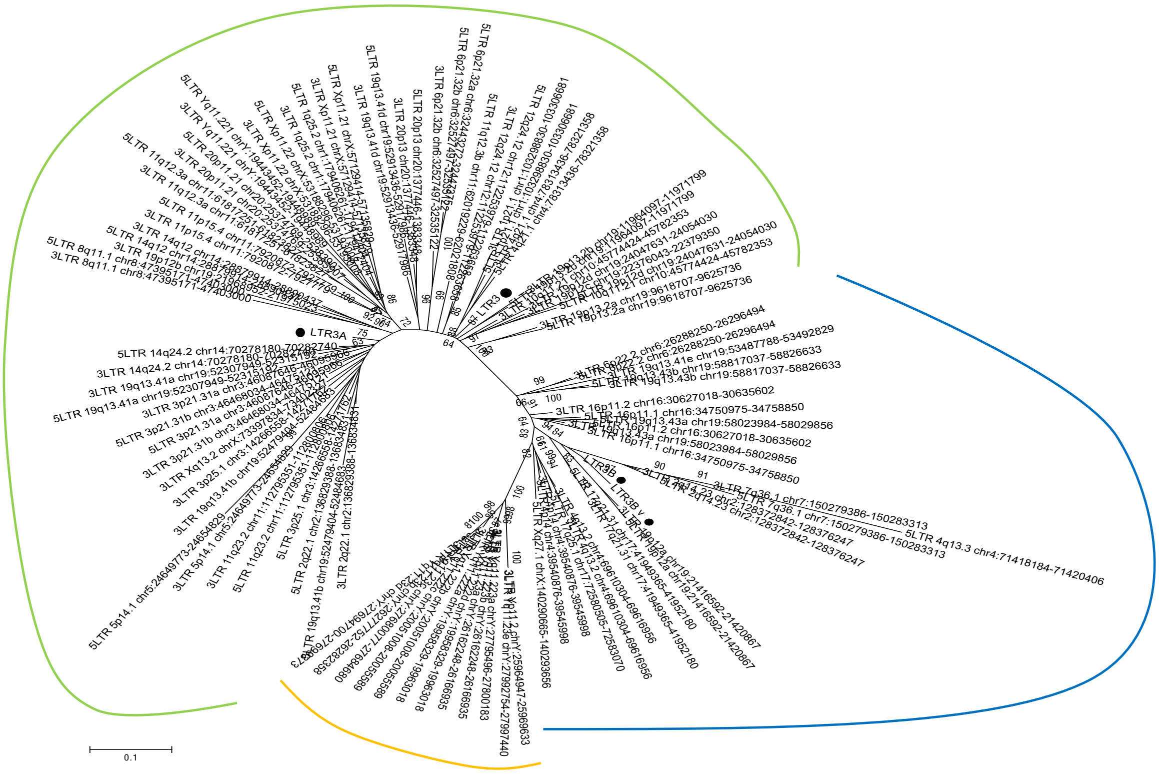

LTR3A
LTR3
LTR3B
LTR3B_v
5LTR 14q24.2 chr14:70278180-70282740
3LTR 14q24.2 chr14:70278180-70282740
3LTR 19q13.41a chr19:52307949-52315192
5LTR 19q13.41a chr19:52307949-52315192
5LTR 8q11.1 chr8:47395171-47403000
3LTR 8q11.1 chr8:47395171-47403000
3LTR 16p11.2 chr16:30627018-30635602
5LTR 4q13.3 chr4:71418184-71420406
5LTR 5p14.1 chr5:24649773-24654829
0.1

Figure 16. Phylogenetic characterization of HML-6 LTRs. The evolutionary distance between the LTRs was shown in a phylogenetic tree including all the HML-6 LTRs and the LTRs Dfam consensus sequences (black dots). The tree showed three different clusters of sequences, one LTR3A and LTR3 (green line), one including LTR3B (light blue line) and one including LTR3B_v elements (orange line).

The presence of LTR3B associated to loci Xq27.1 and 17q25.1, which were assigned to type 2 and type 1b clusters by the NJ and ML analysis respectively, allowed us to identify them as type 1b HML-6 loci. We characterized the LTRs identifying the most conserved structures in Betaretroviral LTRs. The polyadenylation signal was clearly present in position 303-308 as AATAAA box, and we found a putative GT/CT area immediately after this motif. We did not find any evidences of TATA box structure (Figure 15).

Next, we collected the solitary LTRs by using the LTR3A, LTR3, LTR3B and LTR3B_v as consensus sequences for a BLAT search on human assembly hg19, identifying 385 mostly intact LTRs. It is well known that the 5'- and 3'-LTRs of the same provirus are identical at the time of integration [150], and that they independently accumulate random substitutions comparably to the internal proviral sequences and the host genome, allowing to assess the provirus time of integration according to the nucleotide divergence between LTRs. However, due to deletions and rearrangements, in many instances only one (or none) proviral-associated LTR is available, impairing the estimation of the time of insertion. Hence, we recently implemented the calculation of time of insertion with the use of multiple divergence data between individual genic portions and their consensus [18]. Considering a mutation rate in humans of 0,002/nucleotide/million year [151], we estimated the evolutionary age of each HML-6 sequence by calculating nucleotide divergences both between 5'- and 3'-LTRs of each provirus and between 150–350 nucleotide-length portions of *gag*, *pol* and *env* genes and a generated consensus for each subgroup (Figure 17).

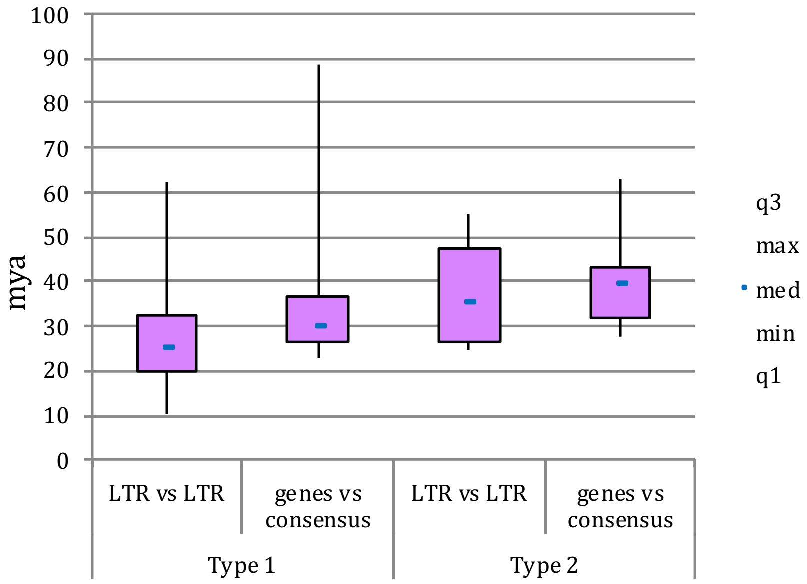


Figure 17. Time of integration of the HML-6 elements. The time of integration was computationally evaluated by using both the divergences between 5'- and 3'-LTRs and between 150–350 nucleotide-length portions of *gag*, *pro*, *pol* and *env* genes and a generated consensus.

While the combination of both LTR-based and consensus-based divergence calculation was possible for only 16 elements, together the two methods allowed estimating the time of integration of overall 54 HML-6 proviruses (82%). Results showed that type 2 elements were probably acquired around 35-40 million years ago, while type 1 elements were probably acquired around 25-30 million years ago, possibly suggesting the existence of two waves of HML-6 viral insertions.

4.6 Genomic context of insertion

The impact of HERVs on the human genome largely depends on their context of integration, since proviral insertions in proximity or within human genes are able to influence their expression, both in sense and anti-sense orientation, depending on i) regulatory activity of LTRs; ii) possible insertion of retroviral splice donor and splice acceptor within the human genes; iii) regulatory activity of anti-sense transcripts [33,152,153]. For this reason, resulting from a negative selecting pressure, HERVs were mainly inserted into intergenic regions, whereas the majority of intragenic insertions occurred in the antisense direction to gene transcription [33,152]. Therefore, we analyzed the context of integration of all the 66 HML-6 elements, attempting to design a map of the elements that could be useful to understand their potential effects on human health through further investigations of the genes involved. We found only 19 sequences (representing about 30% of the HML-6 elements) included into intragenic regions: 11 elements were inserted within coding genes, mainly into introns (9/11), and 8 elements into processed or unprocessed pseudogenes (Table 3).

Table 3. HML-6 genomic context of insertion into human coding and non-coding genes.

Sequence	Type	Gene name	Gene type	Description
1p21.1(+)	1a	RP5-936J12.1 (-)	Known lincRNA	
1q25.2(-)	1a	AXDND1(+)	Known protein coding	Axonemal Dynein Light Chain Domain Containing Protein 1
2q14.23(-)	2	MYO7B(+)	Known protein coding	Myosin VIIB
4p14(-)	1b	UGDH-AS1(+)	Known antisense	UGDH antisense RNA 1
4q21.1(-)	1a	CCNG2(+)	Known protein coding	Cyclin G2
6p21.32b(-)	1a	HLA-DRB6(-)	Known transcribed unprocessed pseudogene	Major Histocompatibility Complex, class II, DR beta 6 (pseudogene)
11p15.4(-)	1a	RP11-494M8.4(-)	Known lincRNA	
11q12.3b(+)	1a	RP11-703H8.9 (-)	Known antisense	
14q12(+)	1a	CTD-2591A6.2(+)	Known lincRNA	
16p11.2(-)	1b	FLJ90415 (-)	Known protein coding	Zinc Finger Protein 689

17q25.1(+)	1b	CD300D(-)
		FLJ31882(+)
			Known protein coding
			Known antisense
				Immune Receptor Expressed On Myeloid Cells 1

19p13.2a(-)	1a	CTC-543D15.3(+)	Known lincRNA	
19p13.2b(-)	1a	DKFZp571K0837(+)	Known protein coding	Zinc Finger Protein 439
19p12c(+)	1a	ZNF676(-)	Known protein coding	Zinc Finger Protein 676
19q13.41a(-)	1a	FPR3(+)	Known protein coding	Formyl peptide receptor 3
19q13.41b(+)	1a	ZNF350(-)	Known protein coding	Zinc Finger Protein 350
19q13.41c(-)	1a	ZNF528(+)	Known protein coding	Zinc Finger Protein 528
19q13.41d(-)	1a	ZNF578 (+)	Known protein coding	Zinc Finger Protein 578
19q13.41e(-)	1b	ZNF702P(-)	Known transcribed processed pseudogene	Zinc Finger Protein 702
19q13.43b(+)	1b	ERVK3-1	Known protein coding	Endogenous Retrovirus group K3 member 1

Interestingly, 7 out of the 11 host genes that include HML-6 within their sequences encode for Zinc-finger proteins and may be involved in transcriptional regulation. In addition, the elements integrated into coding regions showed a prevalent anti-sense orientation with respect to the enclosing genes and seem to be mostly integrated into intronic portions. However, HML-6 loci 16p11.2 and 19q13.41c were inserted into exons (FLJ90415 and ZNF528 respectively) and, in the case of 16p11.2, showed the same sense orientation of the surrounding gene. We hence focused the analysis on locus 16p11.2, finding that the 5'-LTR of this sequence overlapped with the first exon of one of the processed transcript of the ZNF528 gene (Gencode ID ENST00000566673.1), in agreement to the Ensembl and Gencode annotations

[140,144]. Importantly, the ZNF528 gene has been reported to code for a Zinc-finger protein involved in suppressing the apoptosis of hepatocellular carcinoma cells and to be overexpressed in hepatocellular carcinoma (HCC) [154]. We confirm that the sequence 6p21.32b, also known as HERVK3I, is located within the intron 1 of DRB2 and DRB6 pseudogenes, in the MHC region, as already reported by Doxidiase *et al.* [152]. In addition to that, we extended the analysis of the context of insertion to the HML-6 solitary LTRs we detected. Interestingly, we observed that a large number of LTRs were integrated close to or within human genes. Indeed, 284 solo LTRs were included within the sequence of the genes and 97 solo LTRs were integrated within a 10 kb window of distance from the nearest neighbor gene (Figure 18).

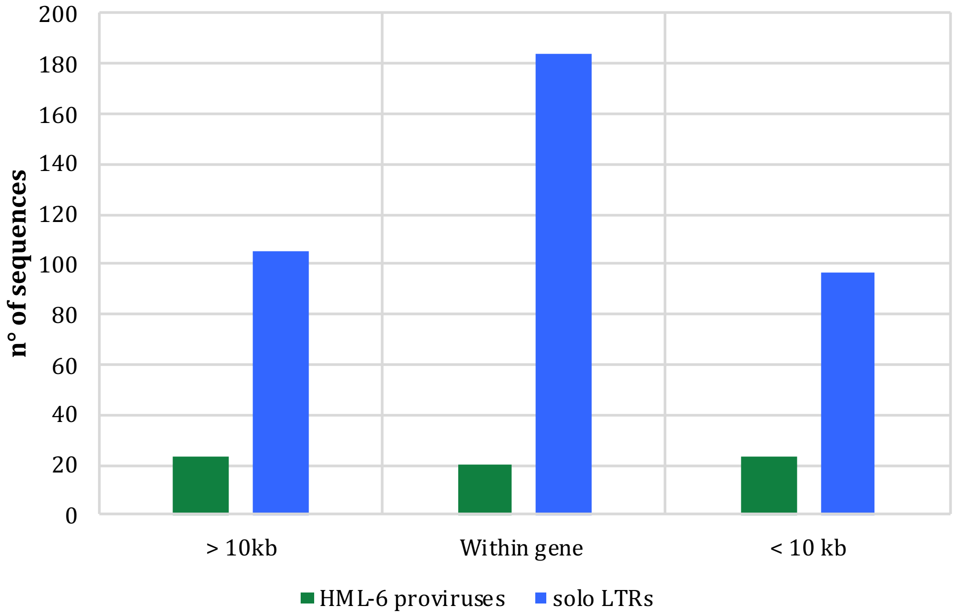


Figure 18. HML-6 proviruses and solo LTRs range of distances from the nearest neighbor gene. The histogram clearly showed a pattern of HML-6 element distribution close to the human genes**.**

4.7 PBS and Betaretroviral structural features

Medstrand *et al.* identified the PBS of the HML-6 sequences to be complementary to lysine tRNA, and consequently included the subgroup into the HERVK clade [98]. Given that such classification was based on a limited number of HML-6 members at the time, we hence aimed to expand that analysis including all the 66 HML-6 sequences collected, to examine possible variations within the subtypes. We found that 38 HML-6 proviruses conserved the PBS regions, 20 of which maintained a well-preserved PBS sequence. As expected, all these PBS were predicted to recognize lysine tRNA (Figure 19).

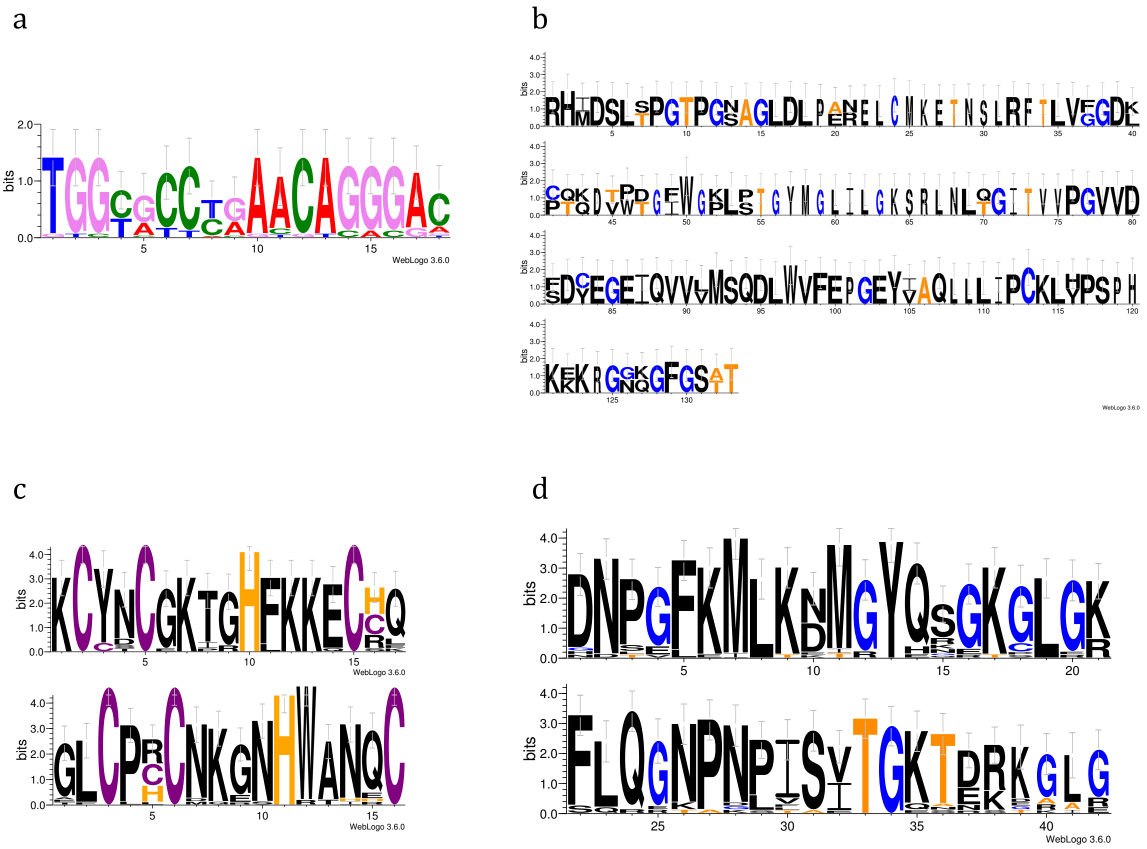


Figure 19. Sequence logos of PBS and beta-retroviral conserved motifs. (a) PBS for lysine nucleotide sequence; (b) Type 1 and type2 Trimeric dUTPase consensus domains; (c) Gag nucleocapsid Zinc fingers and (d) G-patch domain. The height of symbols indicates the relative frequency of each nucleotide or amino acid. Logo created at http://weblogo.berkeley.edu/logo.cgiof.

As further analysis, we searched other typical Beta-retrovirus markers such as two Zinc-fingers in *gag*, and dUTPase and G-patch in *pro* [7,98]. The *gag* Zinc-fingers had a conserved composition (Cys-X2-Cys-X4-His-X4-Cys) and were found in position 1483-1533 (zf-CCHC motif) and 1586-1678 (zf-CCHC_5 motif). We found at least one Zinc-finger motif in > 51% of the HML-6 sequences, and > 13% of the sequences present both zf-CCHC and zf-CCHC_5 motif. In particular, we found 19 loci with a zf-CCHC motif and 24 loci with a zf-CCHC_5 motif. The *pro* dUTPase and G-patch domains have also been detected: we observed a Trimeric dUTPase conserved domain (position 1768-1800) within 20 sequences and a G-patch domain (position 2479-2604) within 12 sequences. Moreover, we also retrieved that the type 1 and type 2 HML-6 dUTPase sequences differed between each other in the first 80 amino acid residues (Figure 19).

4.8 Discussion

The HML-6 group, member of the class II Beta-retrovirus-like HERVs, includes several proviral loci with established transcriptional activity in physiological conditions and an increased transcriptional activity in some human cancers [98,99,101,103]. In particular, two HML-6 transcripts were already shown to contain intact ORFs: i) ERVK3-1 is expressed in various healthy tissues (ENSG00000142396), ii) HERV-K-MEL was shown to encode a small Env peptide in cutaneous and ocular melanoma cells, but not in normal tissues [103]. Nevertheless, due to the absence of a comprehensive description of the HML-6 group at the genomic level, the specific contribution of the individual HML-6 loci to human transcriptome, such as their role in human physiological and pathological conditions, is overall still unclear. In the present thesis chapter, we analyzed in great detail the distribution, genetic composition and phylogeny of all the 66 HML-6 elements retrieved in human genome assembly hg19, providing a complete characterization of the HML-6 group. Overall, based on their chromosomal distribution, HML-6 proviruses showed a random integration pattern, with the only exception of sequences in chromosomes 19, and Y, with a higher number of integration than expected. Such pattern of distribution is in agreement with the ones observed in others HML groups, such as HML-5 and HML-10, and with HERVs in general [7,15,17].

In order to better characterize the group, we analyzed the sequences of the PBS, which has been used for the first classification of the group and was expected to be complementary to lysine tRNA, as reported for the other HML members [7]. Even if, in general, the value of the PBS as a phylogenetic marker is not totally consistent, given the occurrence of alternative PBS types for some HERV groups [18,155], such analysis corroborated the previous findings for the HML-6 elements [98], confirming that they harbor a type-K PBS sequence.

The characterization of the HERVK(HML-6) consensus sequence confirmed a structure resembling the typical proviral genome, with the retroviral genes *gag*, *pro*, *pol* and *env* flanked by 5'- and 3'-LTRs. Worth of note, the structural analysis revealed that 23 HML-6 sequences present a Rec domain, whose presence has been reported here for the first time and has been confirmed through the phylogenetic analysis of Rec putative proteins. The Rec protein, a functional homologue of the retroviral regulatory proteins HIV Rev and HTLV Rex had been initially considered to be present in the sole HML-2 elements [95,96], and HML2 Rec has been shown to interact with the Promyelocytic Leukemia Zinc-Finger protein (PLZF), hence suggesting hypothesis that Rec may contribute to Germ Cell Tumour (GCT) development [156,157]. Recently, the same domain has been also predicted within the sequence of 5 elements belonging to another HML subgroup, HML-10 [17]. Similarly to what observed for

HML-10, no evidence of NP9 protein domain was observed in HML-6, that is hence limited to the sole type 1 HML-2 group [97]. Worth of note, one of the identified Rec domain lies within the ERVK3-1 gene, that has 8 transcripts expressed in several tissues including 6 transcripts predicted to be coding (29, 37). In addition, it is also known that overexpression of HML-2 Rec, in a pluripotent cell line is sufficient inhibit HIV viral infections [60]. This information, together with the previously described HML2 Rec implication in human pathology, suggests the need of further investigations on the role of this domain in HML-6 sequences.

The phylogenetic analysis of the HML-6 internal sequences revealed the presence of two main clusters, that we named type 1 and 2, with type 1 showing an additional internal subdivision in two clusters: type 1a and type 1b. Phylogenetic analysis of the HML-6 Gag amino acid sequences allowed us to confirm the group division into type 1 and type 2 elements. We also confirmed that the HML-6 internal sequences are associated to four types of LTRs: LTR3, LTR3A, LTR3B and LTR3B_v. These LTR types showed differences in sequence length and were grouped into three phylogenetic clusters, one including LTR3A and LTR3, the other including LTR3B and the third including LTR3B_v elements. Indeed, analyses of solo LTRs is necessary for further structural characterizations and the creation of more representative consensus sequences. Interestingly, we observed that type 1a elements were associated only with LTR3 and LTR3A; type 1b sequences with LTR3B; whereas LTR3B_v was only related to type 2. The structural analysis showed the presence of a polyadenylation signal and a putative GT/CT rich region in all the HML-6 LTR types. Target site duplications and T-rich regions are also present in retroviral classes and families of LTRs and it has been proposed as binding site for the cellular factor Sp1 [160,161]. While the GT/CT rich region is present in most HERVs, at the best of our knowledge, its functional role in HML-6 LTRs has not been elucidated yet. Interestingly, as also Benachenhou *et al.* reported, the TATA box was absent, and it is possible to speculate a role of the AATAAA motif as TATA box [161]. Finally, the structural and phylogenetic distances between type 1 and type 2 HML-6 elements seemed to reflect different time of insertion and may indicate two separate integration waves for the two types. These results also suggest that the integration of type 2 HML-6 occurred after the divergence between New World Monkeys and Old World Monkeys, at the time of Catarrhini primate speciation, (about 40 millions years ago). Differently, the integration of type 1 HML-6 elements seems to be specific for Hominoid primates, as it was predicted to be occurred about 30 million years ago, after the divergence between Old World Monkeys and Hominoids. An open question is whether the two subsequent waves of integration for type 1 and type 2 elements can be linked to the fact that the acquisition of a retroviral element into the genome might be favored by the presence of a preexisting endogenous retrovirus through recombination events. Indeed, as it has been recently reported in koalas, the presence of older HERVs facilitates the disruption, and thus endogenization, of a coexisting exogenous species [162].

The analysis of the HML-6 genomic context of insertion and co-localization with functional genes and sequences putatively involved in disease showed that its pattern is comparable to the ones of other HERV elements [152], showing a higher HML-6 presence in intergenic regions, whereas the majority of sequences within intragenic regions resulted integrated in an anti-sense orientation. Interestingly, the HML-6 elements were often integrated within host genes coding for Zinc-finger proteins (7/11), that may be involved in transcriptional regulation. Indeed, we found that sequences 16p11.2, overlaps with a processed transcript of the gene Zinc Finger Protein 689, which is overexpressed in HCC [154]. Anyway, the co-localization with zinc-fingers, which is particularly evident in chromosome 19, may be a consequence of the prevalence of zinc-fingers in this chromosome.

Of particular interest was the finding of a large portion of solo LTRs close to human genes. While we did not perform a complete phylogenetic and structural characterization of all solo HML-6 LTRs, the possible presence of poly-adenylation signals and GT/CT rich regions, observed in the proviral LTRs, may have an influence on the neighbor human gene expression, as already reported in other studies [34,83].

In conclusion, the performed analysis gives complete and updated information on the HML-6 individual loci in the human genome GRCh37/hg19, essential to better understand the genetics of this group, including the possible contribution in physiological and pathological contexts, and its comprehensive transcriptional/ translational analysis.

5. RNA-seq of HERV and MaLR transcriptome analysis in Human Peripheral Blood Mononuclear Cells (PBMCs) in immunity

5.1 RNA-seq analysis reveals HERV and MaLR modulation in PBMCs after in vivo Lipopolysaccharides (LPS) injection

5.1.1 Description of the HERV and MaLR transcriptome in PBMCs

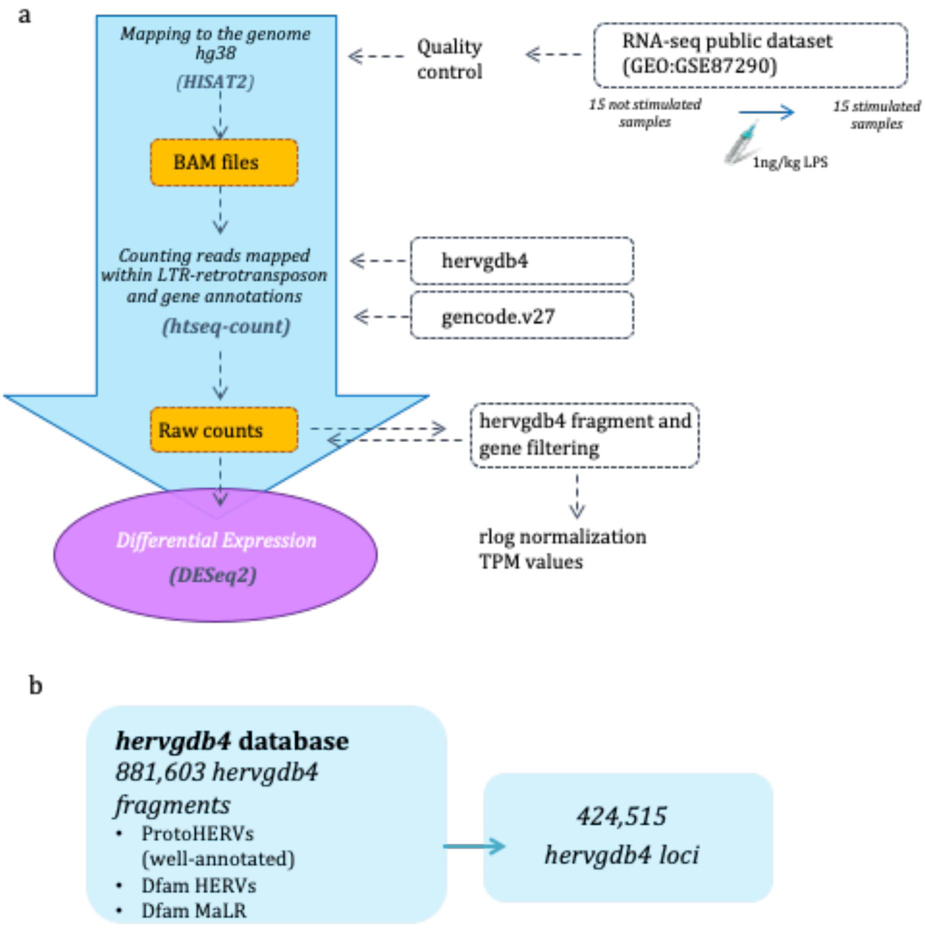
To detect the HERV/MaLR transcriptome in PBMCs, we defined an RNA-seq pipeline to be used on a public RNA-seq dataset (GEO:GSE87290) that included the PBMC transcriptome of 15 healthy volunteers before and 2 hours after *in vivo* stimulation with 1 ng/Kg of LPS (Table 4).

Table 4. Samples and experimental condition from the RNA-seq dataset GSE87290

Run	BioSample	Immune response	Population	Gender	Condition	Sample
SSR4292082	SAMN05806811	High	Afro-American	Female	NS	Sample_01
SSR4292083	SAMN05806810	High	Caucasian	Male	NS	Sample_02
SSR4292084	SAMN05806809	High	Caucasico	Male	NS	Sample_03
SSR4292085	SAMN05806838	High	Caucasico	Male	NS	Sample_04
SSR4292086	SAMN05806837	High	Caucasico	Male	NS	Sample_05
SSR4292087	SAMN05806836	High	Afro-American	Female	NS	Sample_06
SSR4292088	SAMN05806835	Low	Caucasico	Male	NS	Sample_07
SSR4292089	SAMN05806834	Low	Caucasico	Female	NS	Sample_08
SSR4292090	SAMN05806833	Low	Caucasico	Female	NS	Sample_09
SSR4292091	SAMN05806832	High	Caucasico	Female	NS	Sample_10
SSR4292092	SAMN05806831	High	Afro-American	Female	NS	Sample_11
SSR4292093	SAMN05806830	Low	Afro-American	Male	NS	Sample_12
SSR4292094	SAMN05806829	Low	Afro-American	Male	NS	Sample_13
SSR4292095	SAMN05806828	Low	Caucasico	Male	NS	Sample_14
SSR4292096	SAMN05806827	Low	Afro-American	Female	NS	Sample_15
SSR4292097	SAMN05806826	High	Afro-American	Female	LPS stimulation	Sample_01
SSR4292098	SAMN05806825	High	Caucasico	Male	LPS stimulation	Sample_02
SSR4292099	SAMN05806824	High	Caucasico	Male	LPS stimulation	Sample_03
SSR4292100	SAMN05806823	High	Caucasico	Male	LPS stimulation	Sample_04
SSR4292101	SAMN05806822	High	Caucasico	Male	LPS stimulation	Sample_05
SSR4292102	SAMN05806821	High	Afro-American	Female	LPS stimulation	Sample_06
SSR4292103	SAMN05806820	Low	Caucasico	Male	LPS stimulation	Sample_07
SSR4292104	SAMN05806819	Low	Caucasico	Female	LPS stimulation	Sample_08
SSR4292105	SAMN05806818	Low	Caucasico	Female	LPS stimulation	Sample_09
SSR4292106	SAMN05806817	High	Caucasico	Female	LPS stimulation	Sample_10
SSR4292107	SAMN05806816	High	Afro-American	Female	LPS stimulation	Sample_11
SSR4292108	SAMN05806815	High	Afro-American	Male	LPS stimulation	Sample_12
SSR4292109	SAMN05806814	Low	Afro-American	Male	LPS stimulation	Sample_13
SSR4292110	SAMN05806813	Low	Caucasico	Male	LPS stimulation	Sample_14
SSR4292111	SAMN05806812	Low	Afro-American	Female	LPS stimulation	Sample_15

Results showed that we were able to discriminate a total of 424,515 loci included in the *hervgdb4* database, comprising 197,341 HERV loci and 227,174 MaLR loci [10]. Of note, as the *hervgdb4* database has been created as part of the design of Affymetrix HERV-V3 array probes [10], the HERV and MaLR loci are included in the database as 881,603 *hervgdb4* fragments (single genes or functional portion of LTRs) belonging to 424,515 loci (Figure 20) [10]. Moreover, according to the annotation accuracy of the fragments and to their source, the loci were part of two different subsets, i) the highly informative HERV_prototypes and ii) the roughly annotated HERV_Dfam and MaLR_Dfam, both collected from the Dfam database [131]. Analysis showed that 18,633 HERV *hervgdb4* loci and 17,053 MaLR *hervgdb4* loci were expressed in both stimulated and un-stimulated PBMCs samples, for a total of 35,686 loci accounting for ~ 8.4% of the HERVs and MaLRs in the human genome (Figure 20). The majority of expressed HERVs and MaLRs were part of the roughly-annotated HERV_Dfam and MaLR_Dfam subset, for which no information about the groups was available. Among the expressed loci included in the well-annotated HERV_prototypes, 2084 were members of the class I Gamma/Epsilon-like, 527 of the class II Beta-like and 310 of the class III Spuma-like (Figure 21). When considering only the absolute number of expressed loci, just ahead form the HERV-H group, the PRIMA41 group was the one most represented. Instead, when considering the percentage of expressed HERV_prototype loci among the total members of the same group, the most recently integrated group HERV-K(HML-2) showed the highest transcriptional activity, with more than 30% of expressed loci. HML-10, HML-8 and HML-9 groups were also very active, with a proportion of 25.8%, 26.4%, and 25% expressed loci, respectively. In general, the class II Beta-like groups were those showing the greatest percentage of overall activation. The proportions of expressed proviruses as compared to solitary

LTRs within the I, II and III classes were 55%, 61% and 64%, respectively (data not shown).



a
Mapping to the genome hg38
(HISAT2)
BAM files
Counting reads mapped within LTR-retrotransposon and gene annotations
(htseq-count)
Raw counts
Differential Expression
(DESeq2)
Quality control
RNA-seq public dataset (GEO:GSE87290)
15 not stimulated samples
15 stimulated samples
1ng/kg LPS
hervgdb4
gencode.v27
hervgdb4 fragment and gene filtering
rlog normalization
TPM values
b
hervgdb4 database
881,603 hervgdb4 fragments
• ProtoHERVs (well-annotated)
• Dfam HERVs
• Dfam MaLR
424,515 hervgdb4 loci

Expressed elements	hervgdb4 fragments	hervgdb4 loci	HERV proviruses
HERV	31751	18633	676
MaLR	28762	17053	NA
tot	60513	35686	676

Figure 20. Experimental design of Differential Expression analysis. RNA-seq workflow for the identification of modulated HERVs and MalRs (a). The input files used are in blue boxes. The composition of hervgdb4 database is schematized in (b). The amount of expressed hervgdb4 fragments and loci have been obtained by filtering the raw counts and are summarized in the table. The coordinates of the expressed ReTE most intact proviruses have been obtained by using the findOverlaps function from package "Iranges" and the coordinates of expressed hervgdb4 fragments**.**

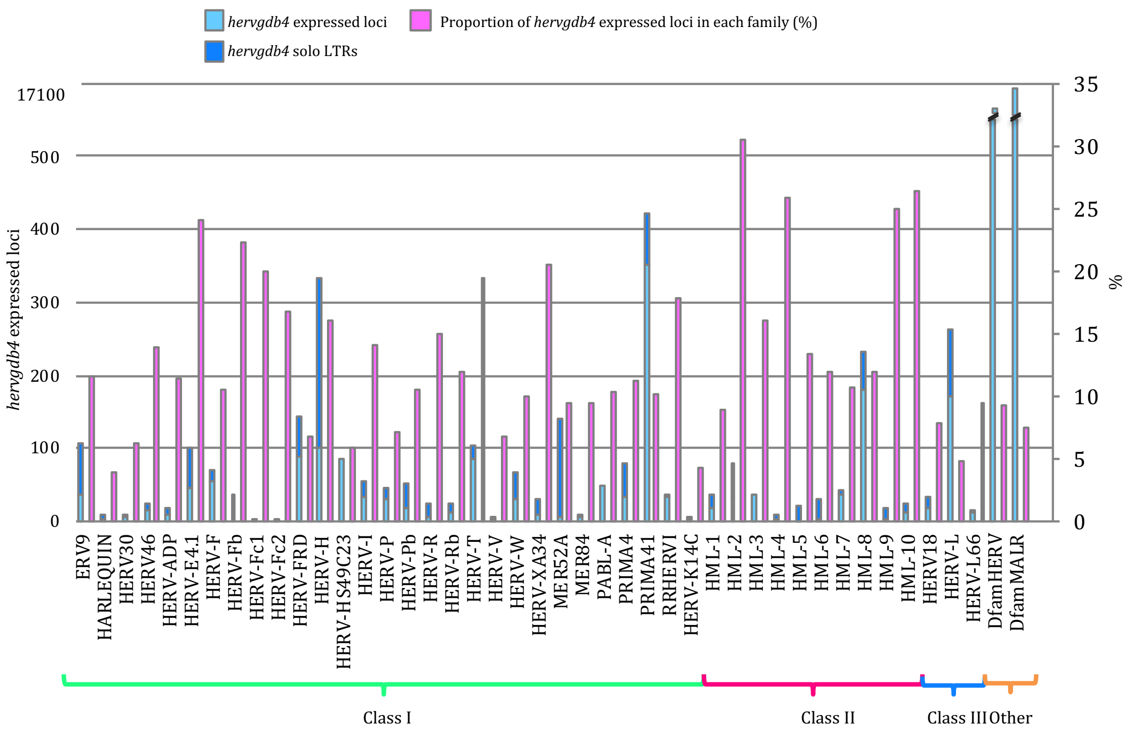


Figure 21. The *hervgdb4* transcriptome in PBMCs. Basal expression of the *hervgdb4* loci in PBMCs. All the expressed elements are grouped by retroviral classes and groups**.**

In order to better define the HERV transcriptome in PBMCs and to assess the link between transcriptional activity and HERV integrity, we decided to analyze the most intact proviruses as identified, classified and characterized in Vargiu *et al.* [7]. Using this dataset, we identified 723 expressed ReTe proviruses, finding also in this case a large proportion of expressed Beta-like elements. HML-4 (6 out of 12 ReTe proviruses) and HERV-K(HML-2) (43 out of 92 ReTe proviruses) were the most active groups (Figure 22). The group with the highest number of expressed ReTe proviruses was the class I Gamma-like HERV-H with 241 active loci, representing 23% of the whole group.

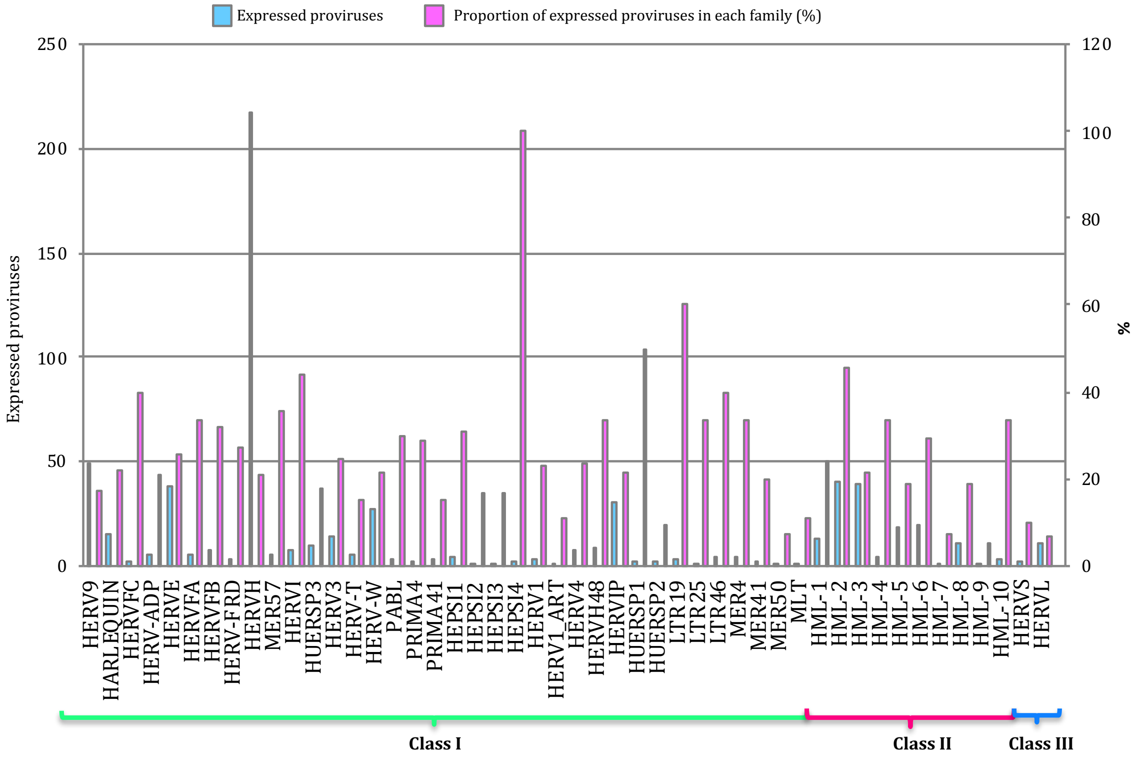


Figure 22. ReTe HERV transcriptome in PBMCs. Basal expression the mostly intact HERV loci reported in Vargiu *et al.* All the expressed elements are grouped by retroviral classes and groups.

5.1.2 Distinct transcriptional patterns induced by LPS stimulation

Next, aiming to assess the variability across the 30 PBMC samples and to detect specific signatures induced by LPS stimulation, we analyzed the expression data of *hervgdb4* fragments using the unsupervised Principal Component Analysis (PCA) (Figure 23). The first Principal Component (PC1) explained the 49% of the variance across samples and the clustering was specifically related to LPS response, showing differences in HERV/MaLR expression between LPS-stimulated and unstimulated samples. Of note, the LPS-stimulated samples 7, 9, 14, and 15 clustered together with the unstimulated ones. These data suggested that LPS response is the principal determinant defining the HERV/MaLR expression inter-sample variability, showing patterns of *hervgdb4* fragments' activation specific for each of the two conditions.

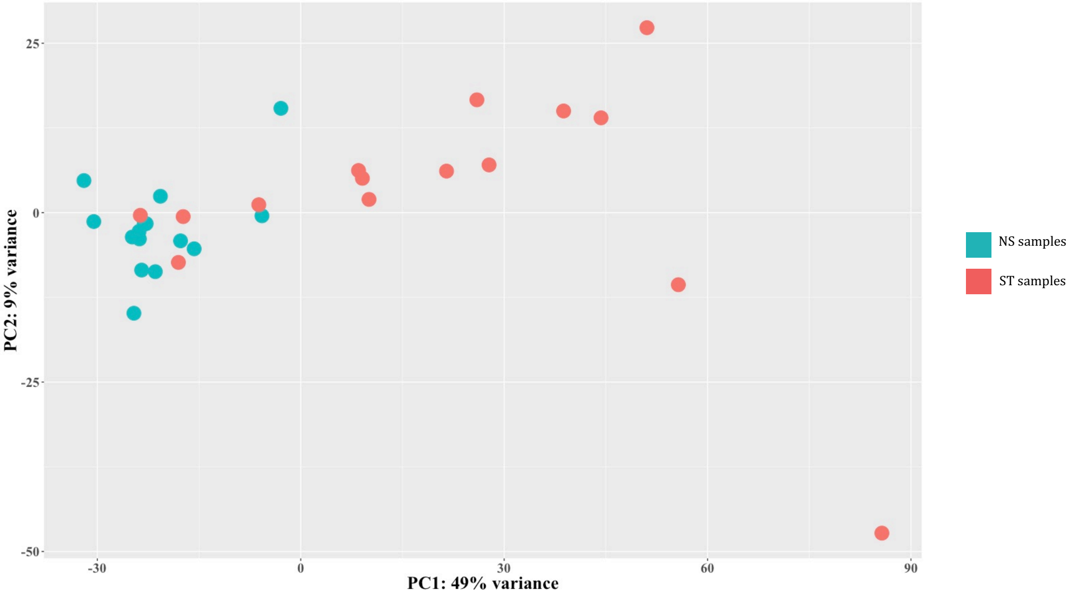


Figure 23. Principal Component Analysis (PCA) of samples. PCA on rlog-normalized *hervgdb4* fragments expression data. It is possible to see the division between not-stimulated and stimulated samples according to the PC1.

We further investigated the transcriptional signatures of the samples by performing hierarchical clustering on the 1,000 *hervgdb4* fragments with the highest mean of read counts across samples (Figure 24). Results showed clear differences in the expression of *hervgdb4* fragments across samples in relation to LPS stimulation. However, both PCA and hierarchical clustering analyses showed that the HERV and MaLR expression profile of samples 7, 9, 14 and 15 after LPS-stimulation was similar to the profiles of the LPS-unstimulated samples. Such behavior was confirmed by the hierarchical clustering performed on the 1000 human genes with the highest mean of reads count among samples, in which the same four samples showed again the typical profiles of the LPS-unstimulated ones (data not shown).

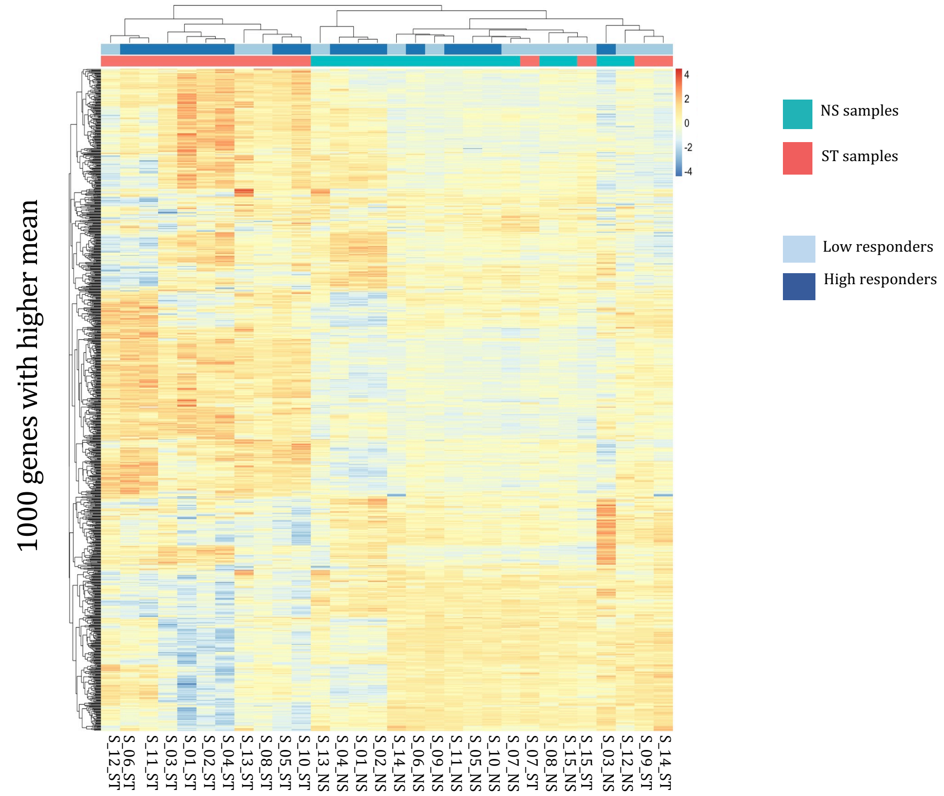


Figure 24. Heatmap of the overall similarity between samples. Hierarchical clustering of the top 1000 *hervgdb4* fragments with the highest average of rlog-normalized counts. The top 1000 *hervgdb4* fragments are in rows and the samples are in columns. rlog-normalized counts are color-scaled from blue (minimum) to red (maximum). Correlation distance measure has been used in clustering columns. Samples are annotated by condition (Not stimulated in aquamarine and LPS-stimulated in red) and inflammatory response (low responders in light blue and high responders in deep blue). The two clusters highlight specific signatures induced by LPS.

This result suggests the presence of transcriptional differences among samples depending on inter-individual variability to immune response, which in turn affect both HERVs/MaLRs and cellular gene expression. Hence, we tried to obtain more information on the interpersonal reaction to LPS investigating the pattern of expression of a subset of 44 genes that have been previously reported to be a specific signature of induced cytokine response [163]. As shown in the heatmap of sample-to-sample euclidean distances, while in the absence of stimulation no evident differences between samples were found for the above 44 genes, after LPS injection two clusters that reflected the traits of the inflammatory response were observed, clearly dividing high-responders from low-responders (Figure 25). Of note, since these 44 genes were shown to be able of deconvoluting complex responses to immune stimulation, we expected that the variability between low- and high-responders would be well-defined. Subsequently, we analyzed the euclidean sample-to-sample distances as defined by the expression of all the *hervgdb4* fragments (Figure 25).

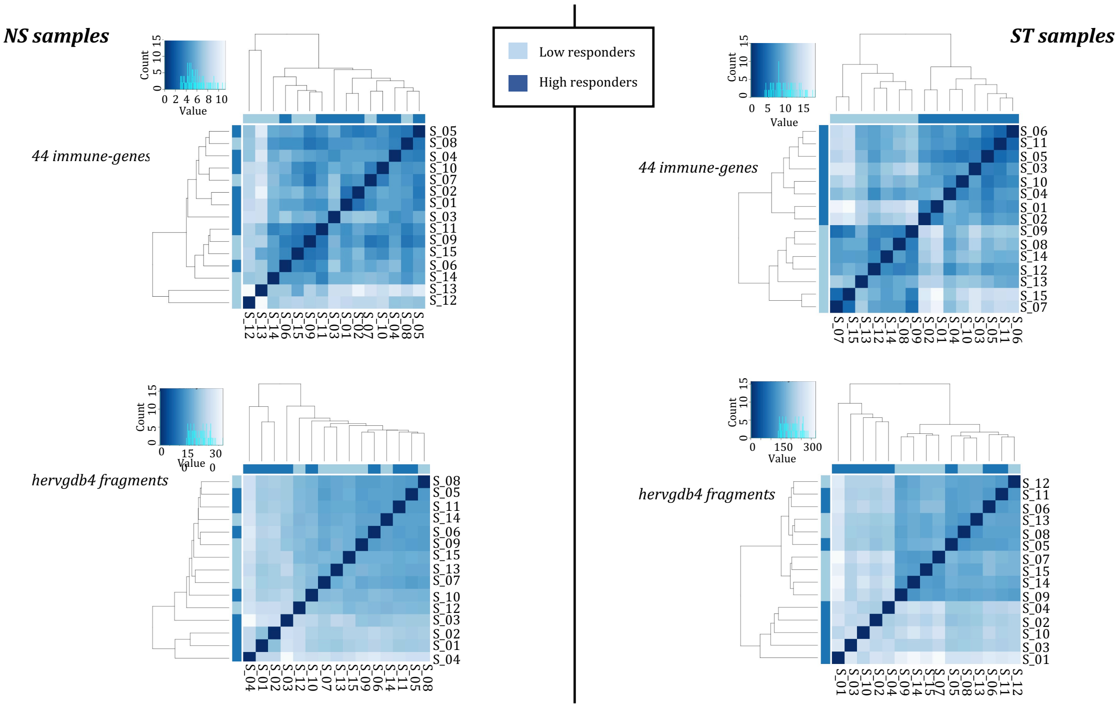


Figure 25. Hierarchical clustering of the Euclidean sample-to-sample distance before (NS samples) and after LPS injection (ST samples). We searched for difference related to the pattern of expression of the 44 genes that captured the diversity of complex innate immune responses (44 immune-genes) and of the *hervgdb4* fragments. The distance values are blue scaled as represented in the color key and histogram legends. The state of inflammatory response of each sample are indicated in light blue (low responders) and deep blue (high responders). High- and low-responders showed different response to inflammation.

Interestingly, as shown for the 44 immunity genes, also for *hervgdb4* fragments no sample clustering was observed in the LPS-unstimulated sample, while in LPS-stimulated samples two clusters were formed, roughly corresponding to the previously identified low- and high-responders to immune stimulation. However, the three low-responder samples clustering with the high-responder ones do not coincide with the four LPS-stimulated samples that showed a pattern of *hervgdb4* fragments expression similar to the LPS-unstimulated samples in figure 24, suggesting that factors other than the severity of inflammatory response may contribute to (rather than interfere with) the inter-individual variability for *hervgdb4* fragments expression.

5.1.3 Differential HERV and MaLR expression in PBMCs

Once assessed that the variability across samples mostly fitted with LPS stimulation, we evaluated the *hervgdb4* fragments for differential expression between the two conditions. After applying a statistical filter (FDR ≤ 0.01 and absolute values of $\log_2$ Fold Change ≥ 1) we identified a total of 6,452 (11% of the total expressed) differentially expressed *hervgdb4* fragments. We represented all the expressed *hervgdb4* fragments in a volcano-plot, where they were indicated as points that spread according to the $\log_2$ Fold Change on the x-axes and to the adjusted p-value on the y-axes (Figure 26). It is worth noting that the great majority of *hervgdb4* fragments were up-regulated, showing a general trend of HERV/MaLR up-regulation in PBMCs after LPS stimulation. In fact, among the 6,452 *hervgdb4* fragments, 5,383 (83%) were up-regulated while 1,069 (17%) were down- regulated after stimulation (Table 5).

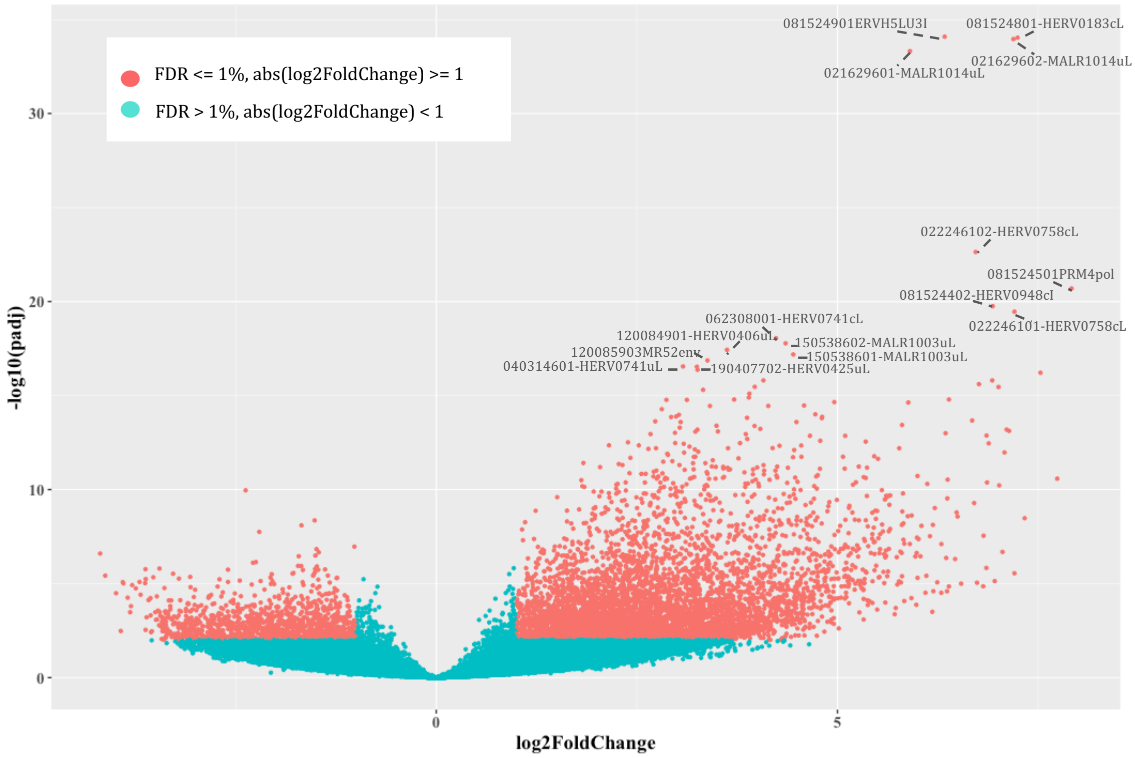


Figure 26. Differential HERV/MaLR expression analysis. Volcano-plot of the differentially expressed *hervgbd4* fragments. Each point represents *hervgbd4* fragments, which spread according to the $\log_2$ fold change (x-axis), and the log10 adjusted p-values (y-axis). Red points are the significantly modulated *hervgbd4* fragments. For the 15 *hervgbd4* fragments with lowest adjusted p-values, the names are indicated.

The mentioned differentially expressed 6,452 *hervgdb4* fragments belonged to 4,607 *hervgdb4* loci, including 3,688 loci (80%) up-regulated, and 919 loci (20%) down-regulated. We then focused on the mostly intact ReTe HERV proviruses, observing that 115 of them were differentially expressed (17% of the ReTe HERV expressed proviruses): 86 were up-regulated while 29 were down-regulated. Out of the 55 HERV differentially over-expressed groups, 6 included only up- regulated proviruses, 13 included both up- and down-regulated proviruses and 5 only down-regulated proviruses. Importantly, we found 23 groups that were constitutively expressed in PBMCs but were not differentially expressed by the stimulation.

Table 5. MaLR/HERVs modulation. Different proportion of modulated, up- and down-regulated *hervgdb4* fragments (a), *hervgdb4* loci (b) and ReTe proviruses (c) in PBMCs after *in vivo* LPS stimulation. *Proportion of expressed elements that are modulated

	Expressed		Modulated	
	Non DE	**DE (%*)**	**Up-modulated**	**Down-modulated**
hervgdb4 fragments	53799	6714 (12%)	5460	1254
hervgdb4 loci	32890	2796 (8%)	1720	1076
HERV proviruses	614	62 (10%)	46	16

5.1.4 Concordant modulation of HERVs and MaLRs and co-localized immunity-related genes

To gain more insights into the HERV/MaLR modulation, we focused on the 15 *hervgdb4* fragments with the highest differential expression according to their adjusted p-value (Table 6, and Figure 26). Importantly, all these 15 highly modulated *hervgdb4* fragments were up-regulated after LPS-stimulation. We used Transcript Per Million (TPM) normalization to quantify the expression levels of the fragments before and after the LPS stimulation (Table 5) and, subsequently, we investigated their context of insertion. Interestingly, we found that, among the 15 most-highly modulated HERVs and MaLRs, 10 were neighbor integrations (within a 10-kb window of distance) of human coding genes. In particular, 6 of them were inside, 3 were downstream and 1 was upstream the colocalized gene. We hence analyzed the modulation of these human genes, observing that all of them were up-regulated as a consequence of LPS stimulation, as summarized in Figure 27.

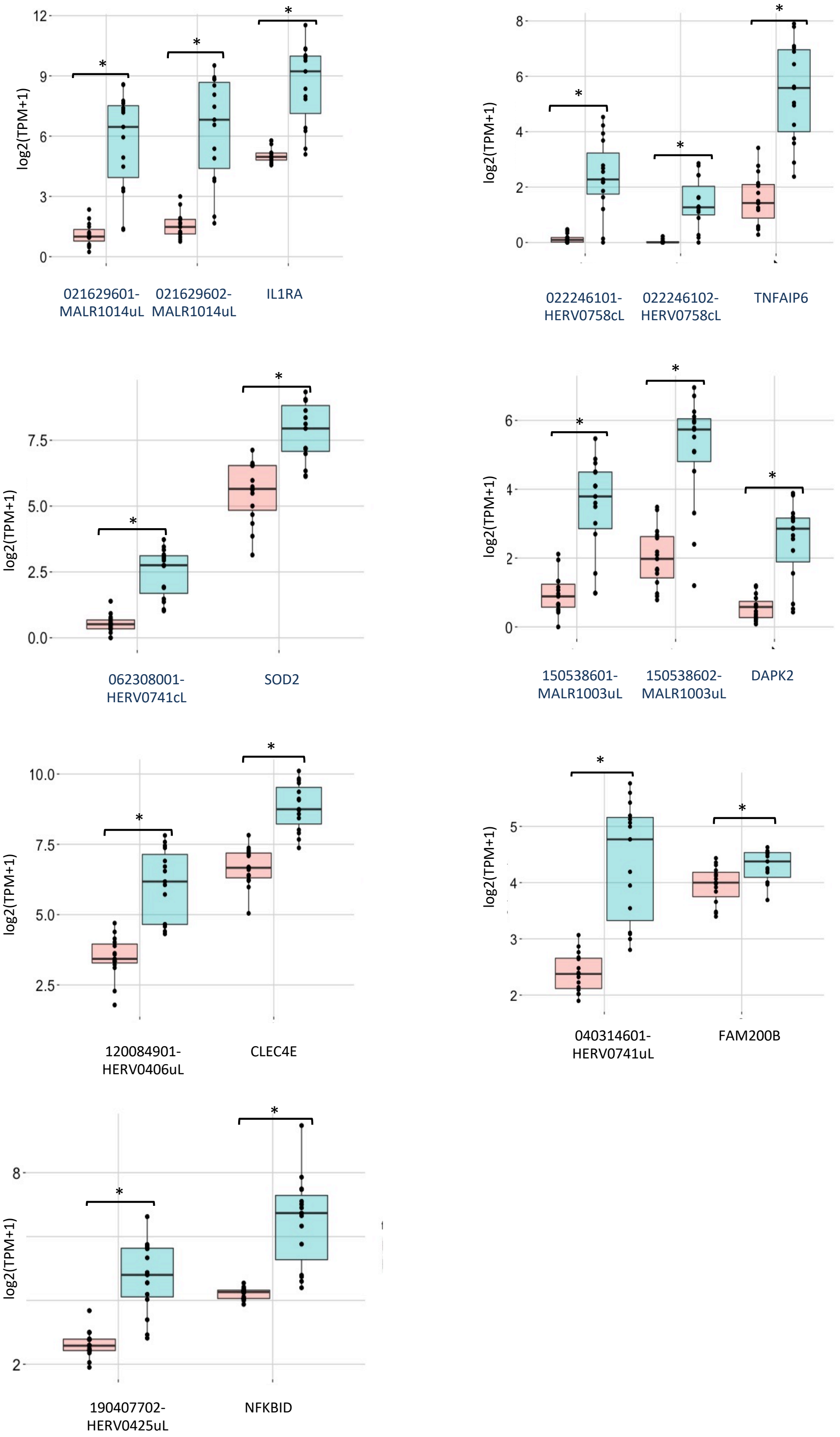


Figure 27. Co-regulation of *hervgbd4* fragments and human genes. Boxplot of the TMP expression values of the *hervgbd4* fragments co-localized with human genes and the neighbor genes. The red boxes indicated values from non-stimulated samples; blue boxes indicated values from stimulated samples. Significant modulations according to the DEseq2 analysis (padj < 0.01) are marked with stars**.**

The fragment with the lowest adjusted p-value (6.99E-36) was 081524901ERVH5LU3I, at coordinates chr8:103002077-103002365. This fragment is the U3 region of a 5'LTR belonging to a HERVH provirus (chr8:103002064-103004587), and its expression levels were increased from an average TPM value of 2.4 to an average TPM value of 124.5 after LPS stimulation.

It is worth to note that, even if this LTR sequence is not co-localized with coding genes, it is integrated into a promoter-flanking region that is affected by copy number variation according to ENSEMBL annotations (data not shown), possibly suggesting a potential transcriptional control role. Interestingly, also fragment 081524801-HERV0183cL, part of a solo LTR (chr8:103001306- 103001748) within the same region as 081524901ERVH5LU3I, increased its average TPM value from 2.4 to 74.2. Fragments 021629602-MALR1014uL and 021629601-MALR1014uL were part of the same solo LTR at coordinates chr2:113131173-113131620. This solo LTR is integrated within the intron of the Interleukin 1 Receptor Antagonist gene (IL1RA), which codes for a protein known to have an anti-inflammatory role [164]. Both gene and solo-LTR significantly increased their expression levels after LPS administration, showing high average TPM values in stimulated samples (Table 6). Similarly, fragments 022246101-HERV0758cL and 022246102-HERV0758cL, part of the same solo LTR at coordinates chr2:113131173-113131620, showed a pattern of up-regulation comparable with their neighbor gene, namely TNF alpha induced protein 6 (TNFAIP6). Thus, in this case, the solo LTR is co-localized with a gene that is involved in immunity, having a known regulatory function [165].

Table 6. Top 15 most DE *hervgdb4* fragments. Description of the 15 most DE hervgdb4 fragments. Same colors indicated fragments of the same locus**.**

Name	*Description*	*Orientation*	*Locus coordinates*	*Host gene*	*Gene coordinates*	*Gene function*	*Position*	*padj*	*log2FC*	*Avg TPM Pre-LPS*	*Avg TPM Post-LPS*
081524901 ERVH5LU3I	LTR (HERV)	NA	chr8:103002064-103004587	NA	NA	NA	NA	8,26E-21	6,3	2,3	99,3
021629602-MALR1014uL	solo LTR (MALR)	-	chr2:113131173-113131620	ENSG00000136689	chr2:113099365-113134016	Interleukin 1 receptor antagonist	Inside	9,23E-29	7,2	2,2	156,7
021629601-MALR1014uL	solo LTR (MALR)	-	chr2:113131173-113131620	ENSG00000136689	chr2:113099365-113134016	Interleukin 1 receptor antagonist	Inside	1,10E-28	7,2	1,4	99,5
081524801-HERV0183cL	solo LTR (HERV)	NA	chr8:103001306-103001748	NA	NA	NA	NA	4,83E-28	5,9	2,4	74,2
022246101-HERV0758cL	solo LTR (HERV)	+	chr2:151381023-151381962	ENSG00000123610	chr2:151357583-151381245	TNF alpha induced protein 6	Downstream	2,33E-17	6,7	0,1	5,6
022246102-HERV0758cL	solo LTR (HERV)	+	chr2:151381023-151381962	ENSG00000123610	chr2:151357583-151381245	TNF alpha induced protein 6	Downstream	2,01E-15	7,9	0,0	2,0
081524501 PRM4pol	INTERNAL (HERV)	NA	chr8:102984015-102986529	NA	NA	NA	NA	1,76E-14	6,9	0,1	3,6
081524402-HERV0948cI	INTERNAL (HERV)	NA	chr8:102980120-102991770	NA	NA	NA	NA	3,48E-14	7,2	0,1	5,5
062308001-HERV0741cL	solo LTR (HERV)	+	chr6:159677828-159678891	ENSG00000112096	chr6:159679064-159693234	Superoxide dismutase 2	Inside	8,89E-13	4,2	0,5	5,5
150538602-MALR1003uL	solo LTR (MALR)	+	chr15:63906995-63907370	ENSG00000035664	chr15:63907036-64040267	Death-associated protein kinase 2	Inside	1,66E-12	4,4	4,0	49,8
120084901-HERV0406uL	solo LTR (HERV)	-	chr12:8537686-8538696	ENSG00000166523	chr12:8533305-8540963	C-type lectin domain family 4 member E	Inside	3,71E-12	3,6	12,0	82,7
150538601-MALR1003uL	solo LTR (MALR)	+	chr15:63906995-63907370	ENSG00000035664	chr15:63907036-64040267	Death-associated protein kinase 2	Inside	6,56E-12	4,4	1,1	14,0
120085903 MR52env	INTERNAL (HERV)	NA	chr12:8566648-8568185	NA	NA	NA	NA	1,36E-11	3,4	5,2	30,7
040314601-HERV0741uL	solo LTR (HERV)	+	chr4:15677257-15678021	ENSG00000237765	chr4:15604539-15681679	Family With Sequence Similarity 200 Member B	Upstream	2,83E-11	3,1	4,4	22,9
190407702-HERV0425uL	solo LTR (HERV)	+	chr19:35887015-35887532	ENSG00000167604	chr19:35887653-35896259	NFKB inhibitor delta	Downstream	2,95E-11	3,2	5,7	28,9

Instead, fragments 081524501PRM4pol and 081524402-HERV0948cI, which are portions of the internal regions in proviral loci chr8:102984015-102986529 and chr8:102980120-102991770, respectively, were found to be intergenic integration. The basal expression levels of both fragments were 0.1 TMP, increasing to 3.6 and 5.5 TPM after stimulation, respectively. Next, we found that fragment 062308001-HERV0741cL, a solo LTR in chr6:159677828-159678891, is integrated within the intron of Superoxide dismutase 2 (SOD2) gene. Of note, fragments 150538602-MALR1003uL and 150538601-MALR1003uL, both part of a solo LTR in chr15:63906995-63907370, showed an average TPM increased from 3.7 and 1.0 to 49.7 and 14.9 after stimulation, respectively, and are integrated within the 3' untranslated region (UTR) of the Death-associated Protein Kinase 2 (DAPK2) gene. Fragment 120084901-HERV0406uL, a solo LTR in locus chr12:8537686-8538696, is integrated inside the C-type lectin domain family 4 member E (CLEC4E) gene and increased its average TPM values from 1.0 to 87.2. The fragments 120085903MR52env (chr12:8566648-8568185), representing an intergenic integration, and 040314601-HERV0741uL (chr4:15677257-15678021), being inserted upstream the Family with Sequence Similarity 200 Member B gene (FAM200B), showed more than 5-folds increase in average TPM after LPS stimulation. Finally, the fragment at coordinates chr19:35887015-35887532, which increased its average TPM values from 5.7 to 28.9, is integrated downstream of the NFKB inhibitor delta gene. For all these HERVs and MaLRs co-localized with human genes, we hypothesized a correlation between their expression levels. We hence visualized the TPM values in scatterplots and measured the Pearson correlation (Figure 28), which allows to quantify possible linear correlation between two variables.

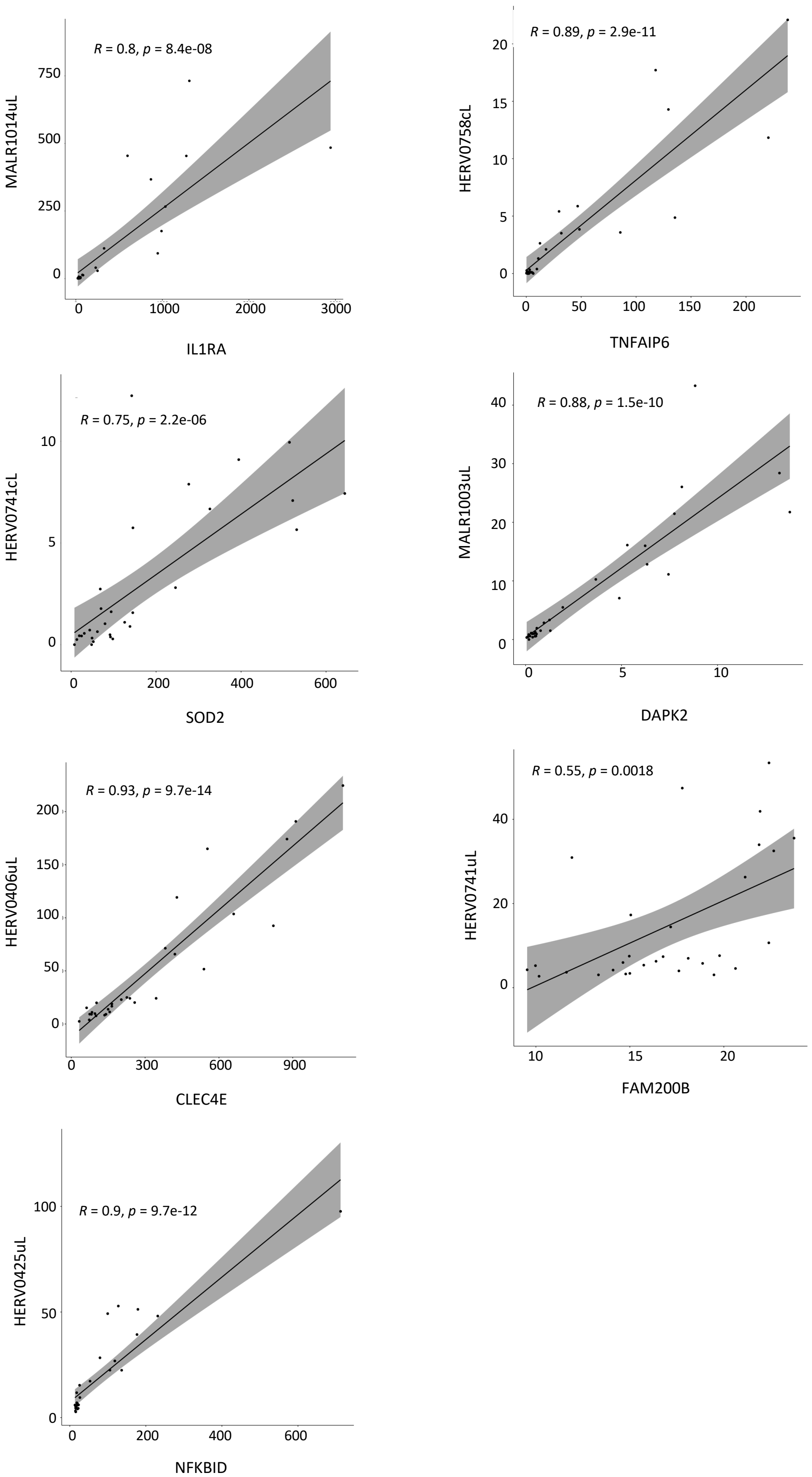


Figure 28. Scatterplots of hervgbd4 fragments and human genes expression. The TMP expression values of the hervgbd4 fragments co-localized with human genes and the neighbor genes are visualized in scatterplots, where the linear regression curve is quantified by Pearson correlation. The correlation coefficients (R) and the p-values (p) of such correlation are indicated in each plot. All the scatterplots showed a positive (R>1) linear correlation between the hervgbd4 fragment and gene TPM values

In all cases, the correlation coefficients (R) where positive, indicating a positive linear correlation between HERVs/MaLRs and co-localized immuno-related genes. In particular, we observed a strong correlation (R>0.7) between all the HERVs/MaLRs and gene pairs, except for that between HERV0741uL (chr4:15677257-15678021) and FAM200B gene that, in any case, was positive (R=0.55).

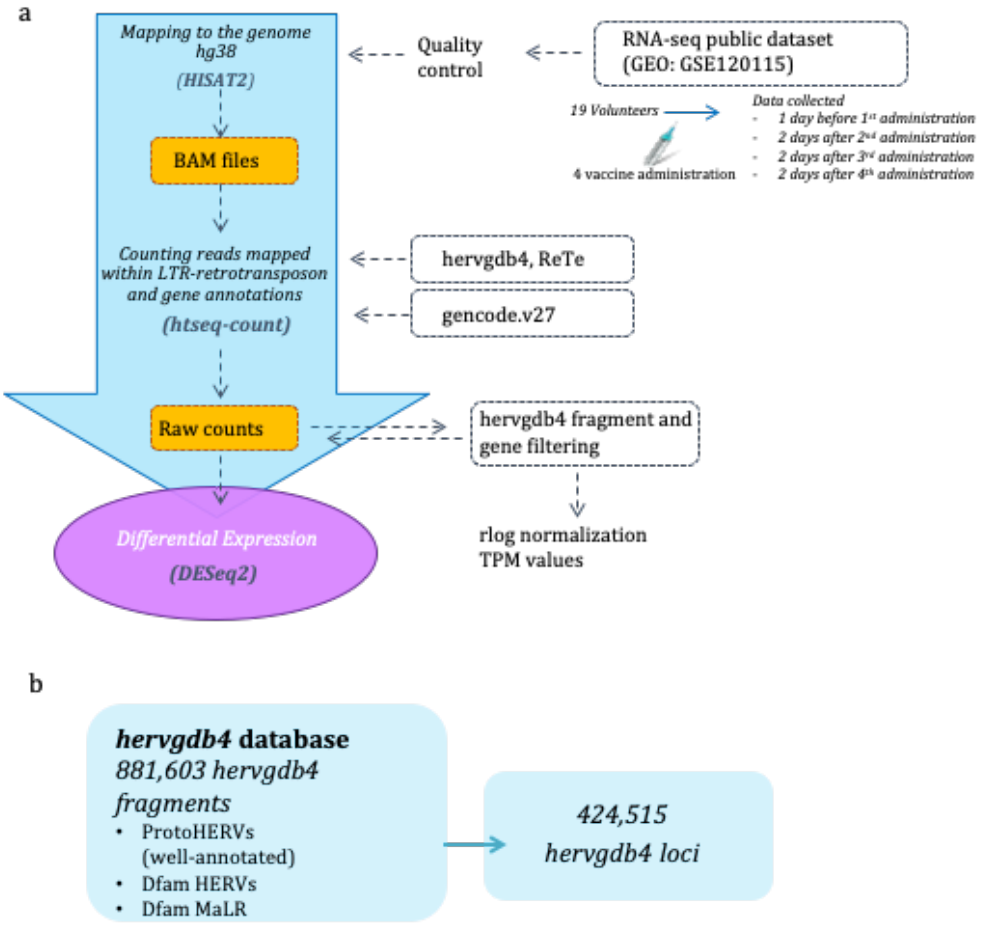
To assess a possible impact of the 10 HERV and MaLR elements on the expression of co-localized cellular genes, we then reconstructed the transcripts associated to their genomic positions in the human genome. In particular, we were interested in the presence of chimeric transcripts, including within their sequences both HERVs/MaLRs and gene portions. Results indicated that no chimeric transcript was present, so that all the 10 HERV and MaLR elements were shown to be transcriptional units different from those of the genes (data not shown).

5.2 RNA-seq analysis of HERV and MaLR and modulation in PBMCs after 4 inactivated virus vaccine administration

5.2.1 Description of the HERV and MaLR transcriptome in PBMCs

We used the same RNA-seq pipeline described above to analyze another RNA-seq dataset (GEO: GSE120115). This dataset includes data from 19 subjects vaccinated against Hantaan virus, which is causative of a hemorrhagic fever with renal syndrome. This vaccine (Hantavax^{TM}) is an inactivated one, and the subjects have been vaccinated with four administration, according to the 0-1-2-13 month schedule [166]. The samples were collected one day before the 1st and two days after the 2nd, the 3rd and the 4th vaccination [166], for a totality of 76 samples (Figure 29, Table 7). Based on neutralizing antibody titers, subjects were classified into non responders, low responders and high responders. Importantly, we chose this specific dataset as the administration of an inactivated vaccine could mimic a viral infection.

We started analyzing the HERV and MaLR transcriptome, with the aim to compare the HERV/MaLR expression patterns in PBMCs among the two datasets (the first one referred to LPS stimulation and the second one referred to vaccination). Data showed that 16,820 HERV *hervgdb4* loci and 15,555 MaLR *hervgdb4* loci were expressed in both vaccinated and pre-vaccinated PBMCs samples (Figure 29), with very similar proportion of expressed loci (7.6%) to those observed for LPS stimulation (8.4%).


a
Mapping to the genome hg38
(HISAT2)
BAM files
Counting reads mapped within LTR-retrotransposon and gene annotations
(htseq-count)
Raw counts
Differential Expression
(DESeq2)
Quality control
RNA-seq public dataset (GEO: GSE120115)
19 Volunteers
4 vaccine administration
Data collected
1 day before 1st administration
2 days after 2nd administration
2 days after 3rd administration
2 days after 4th administration
hervgdb4, ReTe
gencode.v27
hervgdb4 fragment and gene filtering
rlog normalization
TPM values
b
hervgdb4 database
881,603 hervgdb4 fragments
ProtoHERVs (well-annotated)
Dfam HERVs
Dfam MaLR
424,515 hervgdb4 loci

Expressed elements	hervgdb4 fragments	hervgdb4 loci	HERV proviruses
HERV	29535	16820	921
MaLR	26732	15555	NA
tot	56267	32375	921

Figure 29. Experimental design of Differential Expression analysis. RNA-seq workflow for the identification of modulated HERVs and MalRs (a). The input files used are in blue

boxes. The composition of hervgdb4 database is schematized in (b). The amount of expressed hervgdb4 fragments and loci have been obtained by filtering the raw counts and are summarized in the table.

Table 7. Description of samples from dataset GSE120115

Run	Age	BioSample	Gender	Administration	Immune response
SRR7869590	42	SAMN10082551	female	Pre	Low
SRR7869591	42	SAMN10082550	female	2nd	Low
SRR7869592	42	SAMN10082577	female	3rd	Low
SRR7869593	42	SAMN10082576	female	4th	Low
SRR7869594	28	SAMN10082575	female	Pre	High
SRR7869595	28	SAMN10082574	female	2nd	High
SRR7869596	28	SAMN10082573	female	3rd	High
SRR7869597	28	SAMN10082572	female	4th	High
SRR7869598	53	SAMN10082571	female	Pre	None
SRR7869599	53	SAMN10082570	female	2nd	None
SRR7869600	53	SAMN10082569	female	3rd	None
SRR7869601	53	SAMN10082568	female	4th	None
SRR7869602	53	SAMN10082567	male	Pre	None
SRR7869603	53	SAMN10082566	male	2nd	None
SRR7869604	53	SAMN10082565	male	3rd	None
SRR7869605	53	SAMN10082564	male	4th	None
SRR7869606	49	SAMN10082563	female	Pre	Low
SRR7869607	49	SAMN10082562	female	2nd	Low
SRR7869608	49	SAMN10082561	female	3rd	Low
SRR7869609	49	SAMN10082560	female	4th	Low
SRR7869610	52	SAMN10082559	male	Pre	None
SRR7869611	52	SAMN10082558	male	2nd	None
SRR7869612	52	SAMN10082557	male	3rd	None
SRR7869613	52	SAMN10082556	male	4th	None
SRR7869614	50	SAMN10082555	male	Pre	High
SRR7869615	50	SAMN10082554	male	2nd	High
SRR7869616	50	SAMN10082585	male	3rd	High
SRR7869617	50	SAMN10082584	male	4th	High
SRR7869618	48	SAMN10082583	female	Pre	Low
SRR7869619	48	SAMN10082582	female	2nd	Low
SRR7869620	48	SAMN10082553	female	3rd	Low
SRR7869621	48	SAMN10082552	female	4th	Low
SRR7869622	41	SAMN10082604	female	Pre	Low
SRR7869623	41	SAMN10082586	female	2nd	Low
SRR7869624	41	SAMN10082581	female	3rd	Low

SRR7869625	41	SAMN10082580	female	4th	Low
SRR7869626	44	SAMN10082579	female	Pre	None
SRR7869627	44	SAMN10082578	female	2nd	None
SRR7869628	44	SAMN10082625	female	3rd	None
SRR7869629	44	SAMN10082624	female	4th	None
SRR7869630	26	SAMN10082623	female	Pre	High
SRR7869631	26	SAMN10082622	female	2nd	High
SRR7869632	26	SAMN10082621	female	3rd	High
SRR7869633	26	SAMN10082620	female	4th	High
SRR7869634	35	SAMN10082619	female	Pre	High
SRR7869635	35	SAMN10082618	female	2nd	High
SRR7869636	35	SAMN10082617	female	3rd	High
SRR7869637	35	SAMN10082616	female	4th	High
SRR7869638	62	SAMN10082615	male	Pre	Low
SRR7869639	62	SAMN10082614	male	2nd	Low
SRR7869640	62	SAMN10082613	male	3rd	Low
SRR7869641	62	SAMN10082612	male	4th	Low
SRR7869642	38	SAMN10082611	male	Pre	Low
SRR7869643	38	SAMN10082610	male	2nd	Low
SRR7869644	38	SAMN10082609	male	3rd	Low
SRR7869645	38	SAMN10082608	male	4th	Low
SRR7869646	30	SAMN10082607	male	Pre	Low
SRR7869647	30	SAMN10082606	male	2nd	Low
SRR7869648	30	SAMN10082605	male	3rd	Low
SRR7869649	30	SAMN10082603	male	4th	Low
SRR7869650	26	SAMN10082602	male	Pre	Low
SRR7869651	26	SAMN10082601	male	2nd	Low
SRR7869652	26	SAMN10082600	male	3rd	Low
SRR7869653	26	SAMN10082599	male	4th	Low
SRR7869654	34	SAMN10082598	female	Pre	None
SRR7869655	34	SAMN10082597	female	2nd	None
SRR7869656	34	SAMN10082596	female	3rd	None
SRR7869657	34	SAMN10082595	female	4th	None
SRR7869658	31	SAMN10082594	male	Pre	None
SRR7869659	31	SAMN10082593	male	2nd	None
SRR7869660	31	SAMN10082592	male	3rd	None
SRR7869661	31	SAMN10082591	male	4th	None
SRR7869662	26	SAMN10082590	female	Pre	High
SRR7869663	26	SAMN10082589	female	2nd	High
SRR7869664	26	SAMN10082588	female	3rd	High
SRR7869665	26	SAMN10082587	female	4th	High

Hence, we analyzed the HERV_prototypes portion of the *hervgdb4* database, to determine the expression loci distributed among phylogenetic groups (Figure 30). Also in this case, the HERV expression was similar to the one described in data from GSE87290, showing as the expression of the individual HERV groups are rather fixed after microbial infections. When considering the absolute number of expressed loci, the HERV-H and PRIMA41 groups were the most represented. Instead, the HERV-K(HML-2) group showed the highest transcriptional activity, with more than 30% of expressed loci. All groups belonging to class II Beta-like showed high percentage of overall activation.

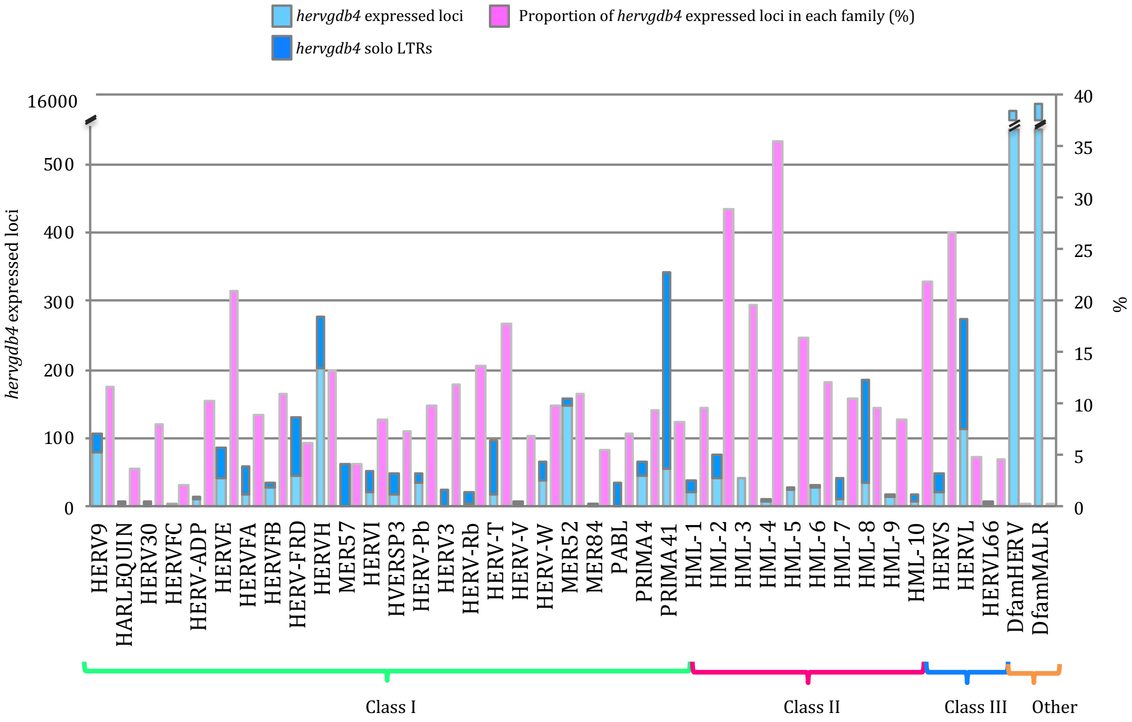


Figure 30. The *hervgdb4* transcriptome in PBMCs. Basal expression of the prototype *hervgdb4* loci. All the expressed elements are grouped by retroviral classes and groups.

We observed similar data also analyzing the most intact HERV proviruses from Vargiu *et al.* [7]. We found expressed 676 ReTe proviruses, and the class II groups showed to be very active, as we also observed in the previous chapter. HERV-K(HML-2) was the most active group, with 47 up to 92 expressed loci (we found 40 expressed loci in basal and LPS-stimulated PBMCs). Similarly, HERVH, with 278 active loci (271 in basal and LPS-stimulated PBMCs), was the group with the highest absolute number of expressed ReTe proviruses. Among the class I (in addition to HERVH), the groups HARLEQUIN, HERV9, HERVE, HERVIP and HERVW were the most active groups, also in this case showing a similar pattern of groups expression to that previously observed (Figure 31).

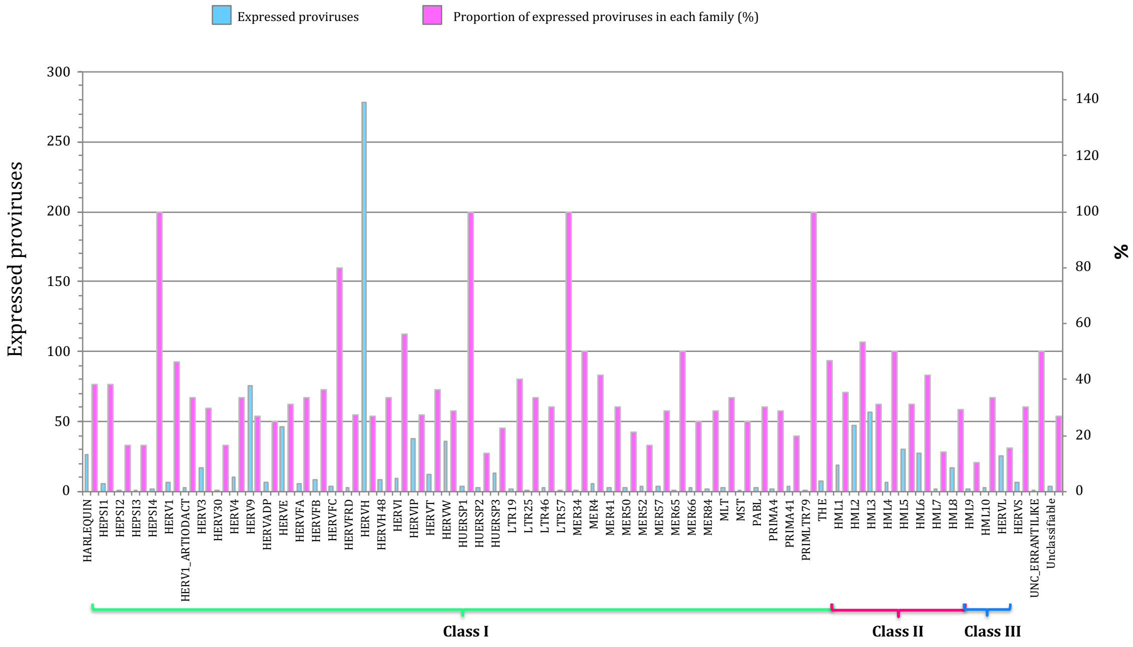


Figure 31. ReTe HERV transcriptome in PBMCs. Basal expression the mostly intact HERV loci reported in Vargiu *et al*. All the expressed elements are grouped by retroviral classes. and groups.

5.2.2 Analysis of transcriptional patterns induced by vaccines

To better understand the balance between innate and adaptive immune response, among the samples we were studying, we analyzed the expression patterns of the subset of 44 genes from Urrutia *et al.*, 2016 previously mentioned. Indeed, these genes give specific signatures of induced cytokine response, linked to innate immunity [163]. The PCA (Figure 32) showed no group cluster of these genes related to the vaccine administration or to the response to the vaccine.

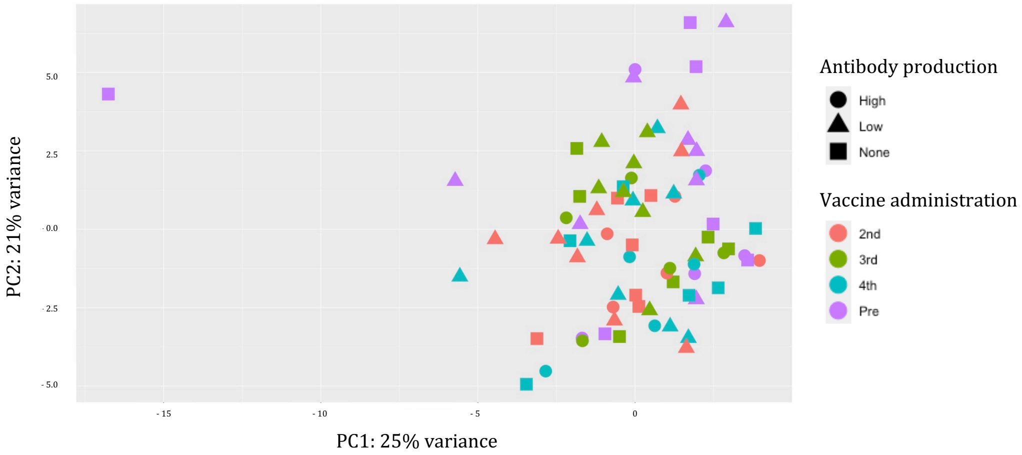


Figure 32. Principal Component Analysis (PCA) of samples according to the expression of 44 genes involved in innate immunity. PCA on rlog-normalized expression data. It is not possible to see any division between vaccinated and not-vaccinated samples, or between samples showing different antibody production.

This result confirmed that the samples, collected two days after the 2nd, 3rd and 4th vaccination, did not show induced cytokine response. Indeed, eventual differences in HERV and MaLR expression are probably guided by mechanisms different from those that act in the first innate immune response. Next, we analyzed the variability among the 76 PBMC samples attributable to the expression of *hervgdb4* fragments, by PCA (Figure 33). Interestingly, the PC1 explained a high proportion of the variance across samples (58%), and 7 samples clustered differently from all the others. This clusters were somehow related to the vaccine administration, as all these 7 samples were pre-vaccinated. Anyway, as only 7 (up to 19) pre-vaccinated samples showed such a distinct expression pattern, HERV and MaLR expression are not representative of the vaccine administration with sufficient precision. Interestingly, the differences in *hervgdb4* expression of the 7 pre-vaccinated samples were not related to their response to vaccine, as the cluster included high- low- and non-responders.

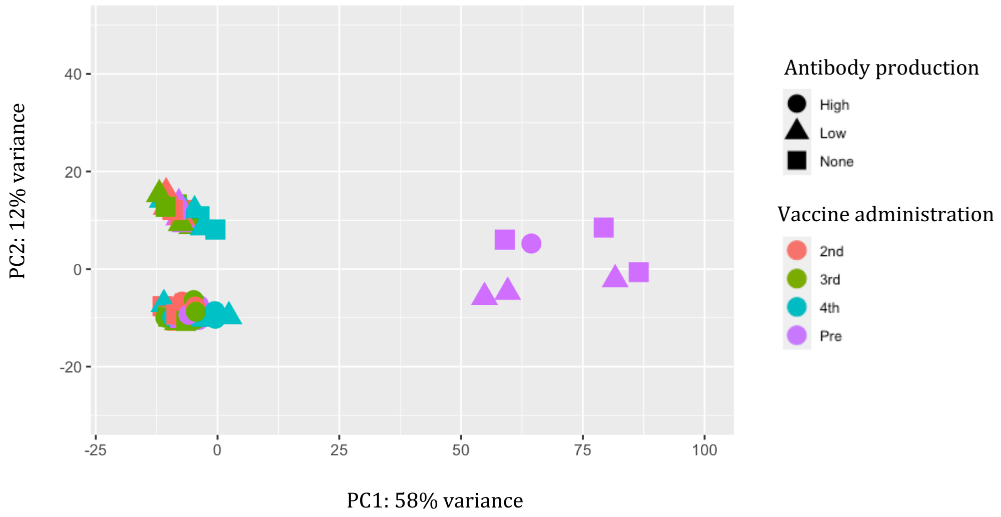


Figure 33. Principal Component Analysis (PCA) of samples. PCA on rlog-normalized *hervgdb4* fragments expression data. The PC1 explains the 58% of the overall variance. It is possible to observe that 7/19 pre-vaccinated samples clustered together. This cluster is not representative of the response to vaccine, as it includes samples belonging to all the three categories, high- low and not- responders. All the other samples seem to show only few differences, explained by the 12% of the variance (PC2).

The other 69 samples also presented few differences, clustering in two groups across the PC2, but these differences were explained by just the 12% of total variance. We obtained similar results by performing hierarchical clustering on the 1,500 *hervgdb4* fragments with the highest standard deviation of reads counts across samples (Figure 34). Results showed the same 7 pre-vaccinated samples having clear distinct patterns of *hervgdb4* fragments expression. Such behavior was also confirmed by the individual hierarchical clustering on pre-vaccinated samples and 2nd–, 3rd– and 4th-vaccine administered samples (Figure 35). Moreover, a further hierarchical clustering on the 1,500 human genes with the highest standard deviation of reads across samples showed again the typical clustering of samples (data not shown).

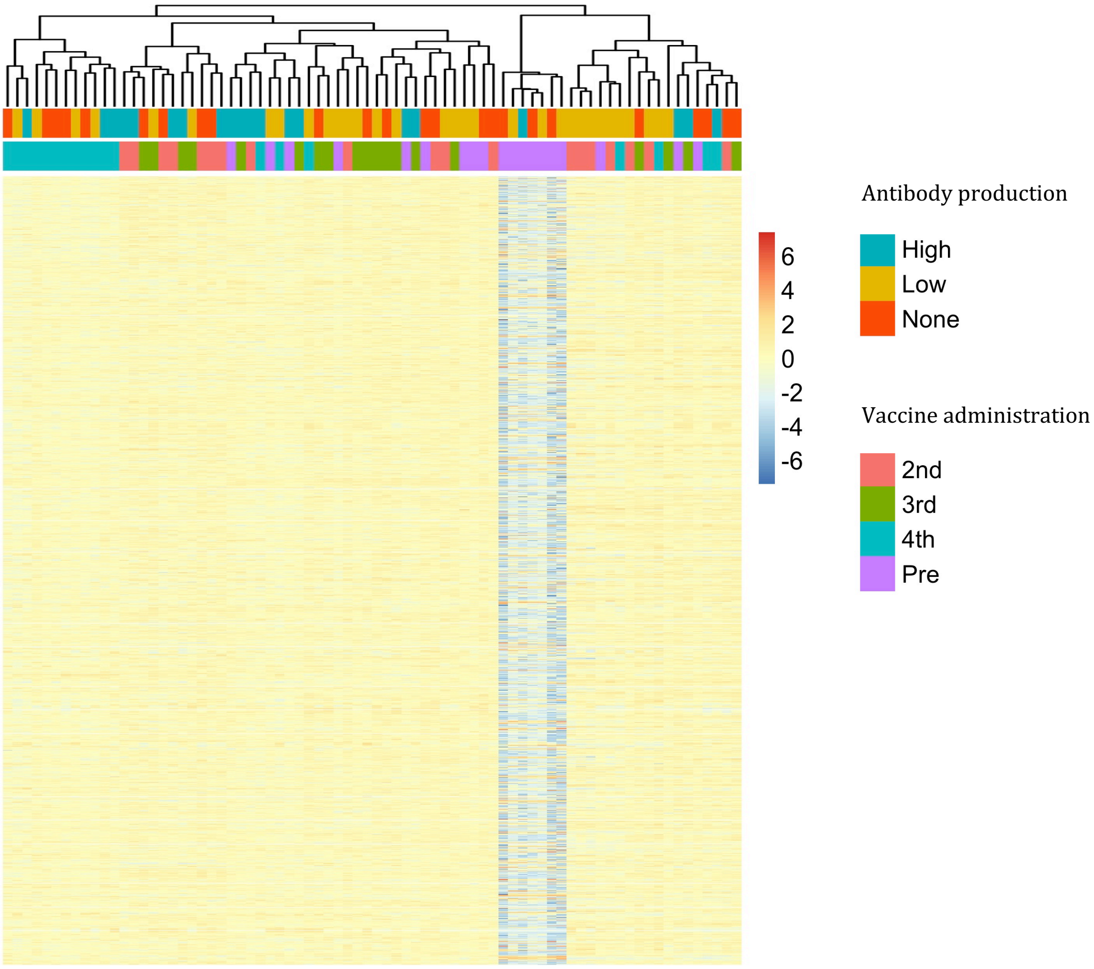


Figure 34. Heatmap of the overall similarity between samples. Hierarchical clustering of the top 1500 *hervgdb4* fragments with the highest standard deviation of rlog-normalized counts. The top 1500 *hervgdb4* fragments are in rows and the samples are in columns. rlog-normalized counts are color-scaled from blue (minimum) to red (maximum). Correlation distance measure has been used in clustering columns. Samples are annotated by antibody production and vaccine administration. Also in this case, 7/19 samples showed different expression patterns from the others.

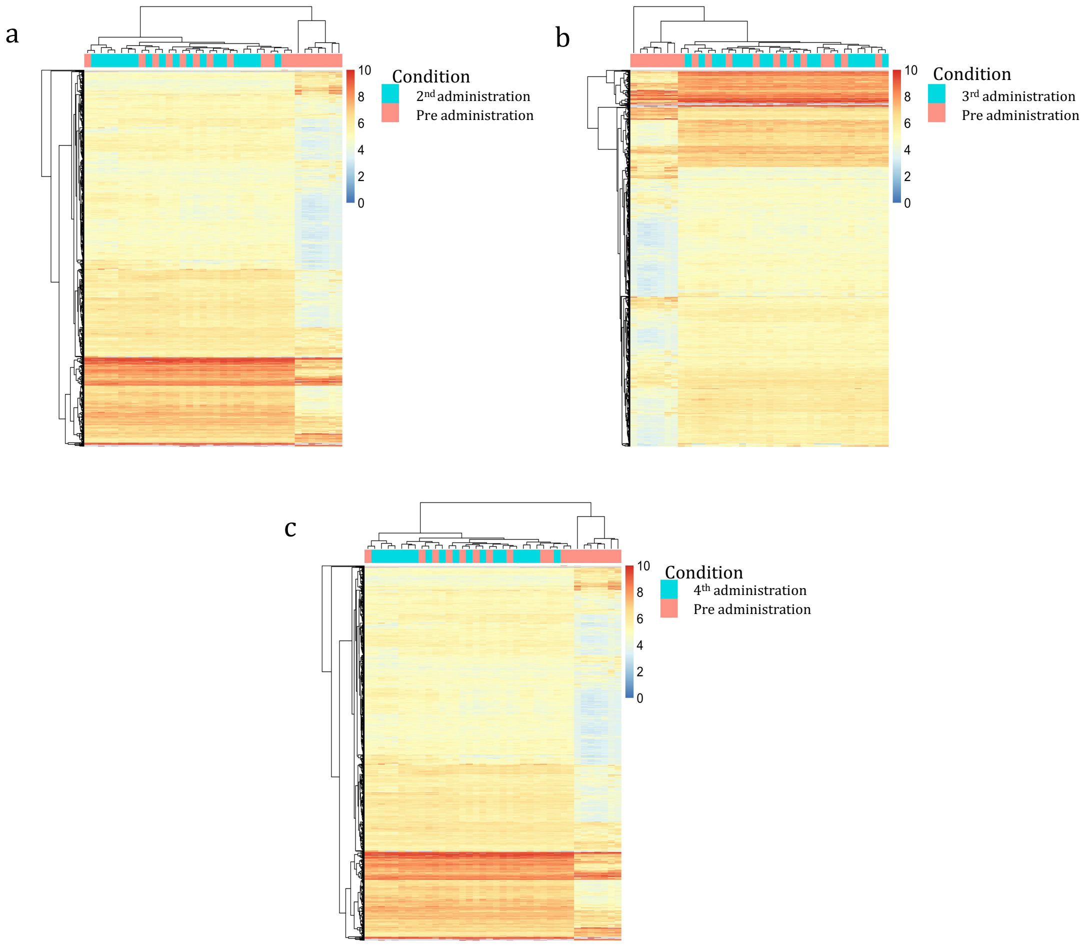


Figure 35. Heatmap of the overall similarity between pre-vaccinated samples and samples after the 2nd , 3rd and 4th vaccine administration. Hierarchical clustering of the top 1500 *hervgdb4* fragments with the highest standard deviation of rlog-normalized counts. The top 1500 *hervgdb4* fragments are in rows and the samples are in columns. rlog-normalized counts are color-scaled from blue (minimum) to red (maximum). Correlation distance measure has been used in clustering columns. Samples are annotated by vaccine administration. In all the three heatmaps, showing pre-vaccinated samples vs 2nd administration (a), 3rd administration (b) and 4th administration (c), 7/19 samples had different expression patterns from the others after all the vaccine administration.

5.2.3 Differential HERV and MaLR expression after vaccine administration

We evaluated the *hervgdb4* fragments for differential expression for three different combination of conditions: i) pre-vaccination and 2nd administration, ii) pre-vaccination and 3rd administration and iii) pre-vaccination and 4th administration. We applied a statistical filter (FDR ≤ 0.01 and absolute values of $\log_2$ Fold Change ≥ 1) to identify the modulated elements, that we represented in a volcano-plot, where they were indicated as red points (Figure 36). Similarly to what we observed after LPS stimulation, the great majority of *hervgdb4* fragments were positively modulated, showing a general trend of HERV/MaLR up-regulation after each vaccine administration. Anyway, the number of modulated elements was very reduced.

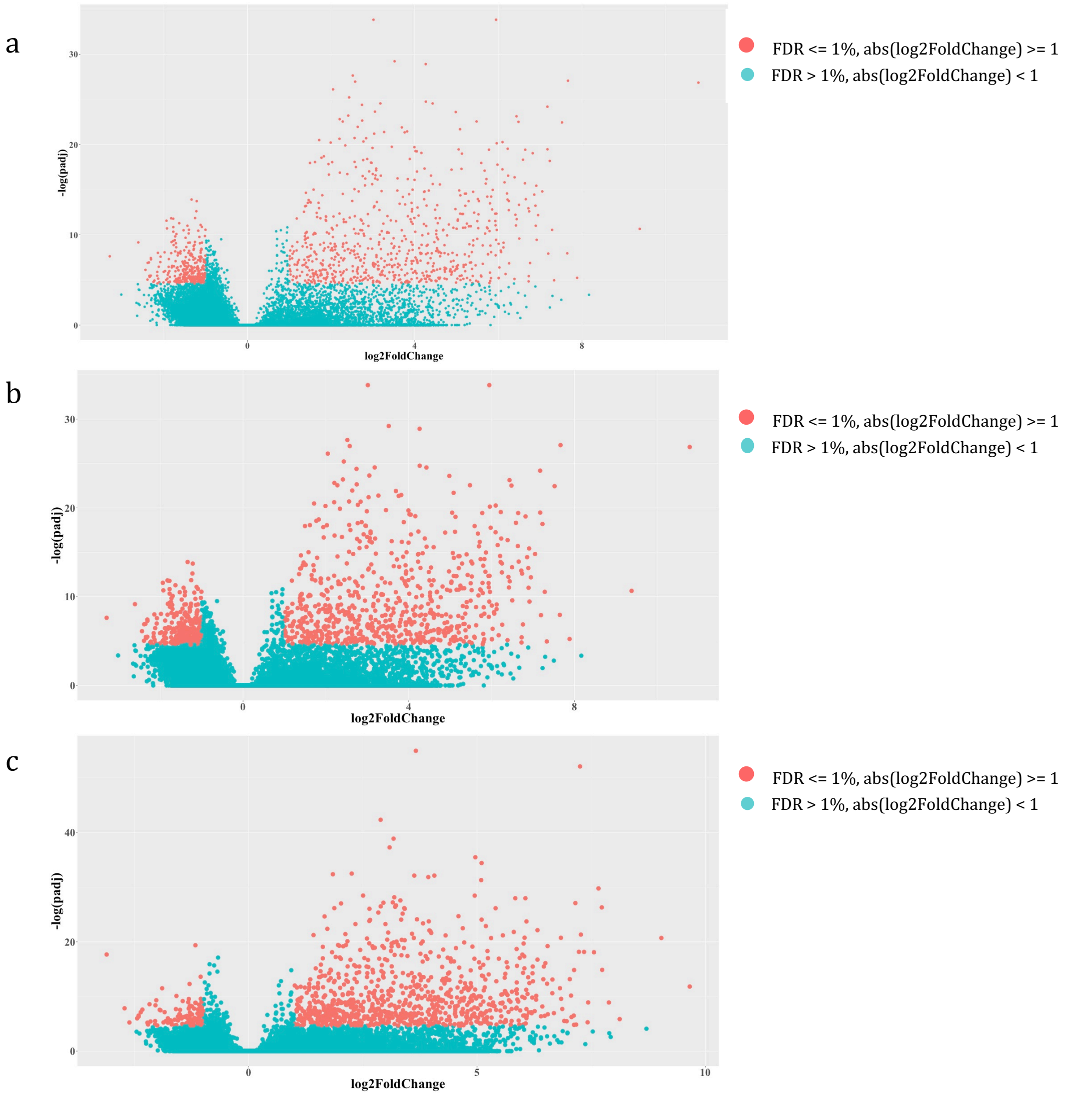


Figure 36. Differential HERV/MaLR expression after vaccine administration. Volcano-plot of the differentially expressed *hervgbd4* fragments after after 2nd , 3rd and 4th vaccine administration are respectively in (a), (b) and (c). Each point represents *hervgbd4* fragments, which spread according to the $\log_2$ fold change (x-axis), and the log10 adjusted p-values (y-axis). Red points are the significantly modulated *hervgbd4* fragments.

1,032 *hervgdb4* fragments (3.4% of the total expressed) were differentially expressed after the 2nd administration, 732 (2.4% of the total expressed) were differentially expressed after the administration and 1,038 3rd (3.5% of the total expressed) were differentially expressed after the 4th administration (Table 5). When considering the *hervgdb4* loci, we found that 608 loci were modulated after the 2nd, 396 after the 3rd and 576 after the 4th administration. Finally, we found 23, 17 and 62 mostly intact ReTe proviruses differentially expressed after the 2nd, 3rd and 4th administration, respectively.

Table 8. MaLR/HERVs modulation. Different proportion of modulated, up- and down-regulated *hervgdb4* fragments (a), *hervgdb4* loci (b) and ReTe proviruses (c) in PBMCs after *2nd* (a), 3rd (b) and 4th (c) vaccine administration.

a

	Expressed		Modulated	
	Non DE	**DE (%*)**	**Up-modulated**	**Down-modulated**
hervgdb4 fragments	29535	1032 (3.4%)	718	314
hervgdb4 loci	32375	608 (1.9%)	416	193
HERV proviruses	921	23 (2.4%)	16	7

b

	Expressed		Modulated	
	Non DE	**DE (%*)**	**Up-modulated**	**Down-modulated**
hervgdb4 fragments	29535	732 (2.4%)	707	25
hervgdb4 loci	26732	396 (1.5%)	381	15
HERV proviruses	56267	17 (0.03%)	16	1

c

	Expressed		Modulated	
	Non DE	**DE (%*)**	**Up-modulated**	**Down-modulated**
hervgdb4 fragments	29535	1038 (3.5%)	910	128
hervgdb4 loci	26732	576 (2.15%)	513	63
HERV proviruses	56267	62 (0.04%)	25	2

Next, we tried to understand if the differentially expressed elements would be specific signature for the various steps of the response to the vaccine, searching for *hervgdb4* fragments and ReTe proviruses that were modulated after all the administrations (Figure 37). The Venn diagram with the intersections of the modulated elements in the three considered conditions showed that a large number of *hervgdb4* fragments and ReTe proviruses were modulated after more than one administration.

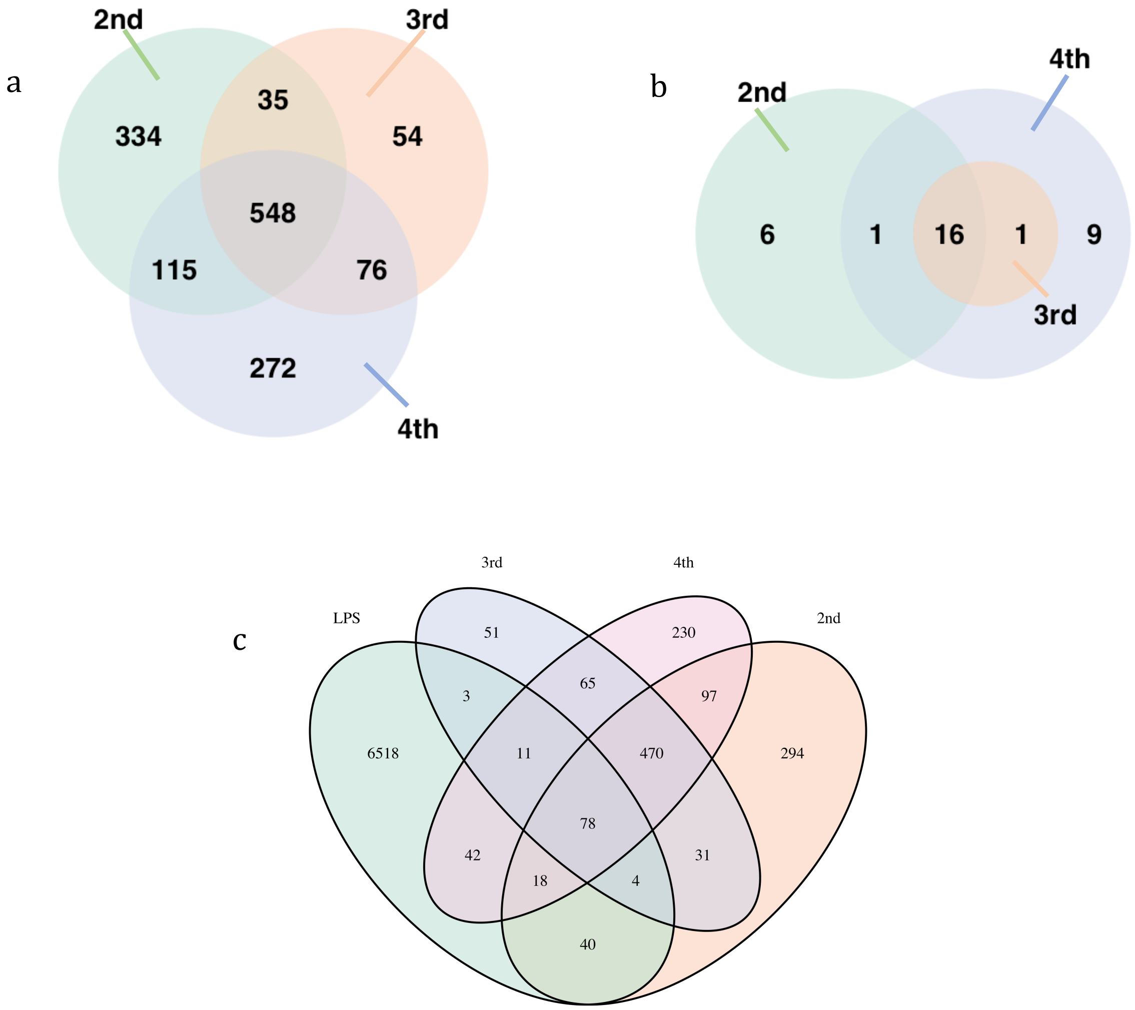


Figure 37. Venn diagrams of the differentially expressed elements. The first diagram (a) shows the intersections of *hervgdb4* fragments modulated after all the vaccine administrations. The second diagram (b) shows the intersections of the ReTe proviruses modulated after the vaccines. Finally, the third diagram (c) shows the intersections of *hervgdb4* fragments modulated after the vaccine administrations (in this work considered as indicative of adaptative immune response) and after LPS stimulation (in this work considered as indicative of innate immune response).

For example, 548 *hervgdb4* fragments and 16 ReTe proviruses were modulated after all the administration. In the case of ReTe proviruses, all the loci modulated after the 3rd administration are also modulated after the 4th. These results suggest that all the vaccine administration trigger similar patterns of HERVs and MaLRs. Among the 16 ReTe proviruses modulated after all the vaccine administration, 5 belonged to the HERVH group (Table 9). Almost all the elements maintained the great majority of the retroviral gene domains. Anyway, further analyses are required to reconstruct the transcripts and better understand which of the genes are expressed.

Table 9. Description of the 16 ReTe proviruses modulated after all the vaccine administration

Chr	Start	End	Strand	Length	ID	Subgenes	Group
chr3	107564215	107572787	-	8572	1058	5LTR PBS MA CA NC Prot IN PPT	HERVH
chr2	69789472	69799355	-	9883	565	5LTR PBS MA CA NC Prot RT IN TM PPT 3LTR	HUERSP3
chr10	18570092	18577466	+	7374	3200	CA NC RT IN TM	HERVIP
chr5	82267546	82273706	-	6160	1892	5LTR PBS CA NC RT IN SU TM PPT 3LTR	HARLEQUIN
chr19	36149712	36161023	-	11311	4713	5LTR PBS MA CA NC Prot RT IN TM PPT	HERVH
chr3	193599956	193613333	-	13377	1278	5LTR PBS CA NC Prot RT IN TM 3LTR	HEPSI1
chr11	58769831	58777331	+	7500	3503	MA NC RT IN SU TM PPT 3LTR	HERV1
chr7	43853008	43866752	-	13744	2476	MA NC Prot RT SU TM PPT	HML3
chr6	148639772	148645510	+	5738	2371	5LTR PBS MA CANC Prot RT 3LTR	HERVH
chr17	11971744	11978102	+	6358	4426	5LTR PBS MA CA NC Prot RT IN	HERVH
chr22	16611312	16616782	+	5470	6262	5LTR PBS MA CA NC Prot IN SU	HERVH
chr5	70512460	70531584	+	19124	1874	5LTR CA NC RT IN TM PPT 3LTR	THE
chr1	155650288	155659631	-	9343	6072	5LTR PBS MA CA NC RT IN SU TM 3LTR	HERV4
chr4	139442392	139449817	+	7425	1638	5LTR PBS RT IN TM PPT 3LTR	HERVL
chr4	53236811	53255667	-	18856	1405	5LTR IN TM PPT 3LTR	HML2
chr4	25238665	25247155	-	8490	1350	5LTR CA NC Prot RT IN 3LTR	HERV9

Then, we checked for intersection with *hervgdb4* fragments modulated after LPS stimulation. In this case, only 196 *hervgdb4* fragments were modulated after both LPS and vaccine injection, highlighting how the HERV and MaLR modulation is different between innate immunity activation and response to vaccine. In the case of ReTe proviruses, only 4 were modulated after both vaccine administration and LPS stimulation. In particular, 2 of them (ID: 9883 and 5470, loci chr2:69789472-69799355 and chr22:16611312-16616782) were modulated after the injection of LPS and after the 2nd vaccine administration, while the proviruses with ID 5731 and 7374, respectively in loci chr8:97175022-97180753 and chr17:77167827-77175201, were modulated after the injection of LPS and all vaccine administration (Table 10). Moreover, among these 4 ReTe proviruses, 3 (chr2:69789472-69799355, chr22:16611312-16616782 and chr17:77167827-77175201) were co-localized with human genes but none was co-localized with genes involved in immunity.

Also among the 15 *hervgdb4* fragments most modulated after the 2nd, 3rd and 4th vaccine administration we did not find any *hervgdb4* fragment co-localized immunity related genes, indeed, all the fragments were intergenic or intragenic of genes that were basally expressed.

Table 10. Description of the 4 ReTe proviruses modulated after both injection of LPS and all vaccine administration

Chr	Start	End	Strand	Length	ID	Subgenes	Group
chr2	69789472	69799355	-	565	9883	5LTR PBS MA CA NC Prot RT IN TM PPT 3LTR	HUERSP3
chr22	16611312	16616782	+	6262	5470	5LTR PBS MA CA NC Prot IN SU	HERVH
chr8	97175022	97180753	+	2909	5731	5LTR PBS MA CA NC Prot RT IN 3LTR	HERVH
chr17	77167827	77175201	-	4488	7374	5LTR PBS MA CA NC Prot IN TM PPT 3LTR	HERVIP

5.3 Discussion

We used an RNA-seq approach and the *hervgdb4* database [10] to obtain an overview of the specific HERV and MaLR transcriptome in PBMCs.

Previously, the same database has been used by Mommert *et al.* [111] to identify expressed and modulated loci in an *ex vivo* system of LPS stimulation and endotoxin tolerance, through microarray analyses [111]. In this regard, the percentage of expressed elements in this *in vivo* model (about 7% of *hervgdb4* fragments and 8% of *hervgdb4* loci expressed) was slightly higher than the one measured in *in vitro* experiments (5.6% of *hervgdb4* fragments). Among all HERV groups, class II members appeared to be the most active ones, with HERV-K(HML-2) being the most expressed group in PBMCs from both the RNA-seq datasets we analyzed. This group is also one of the most investigated, due to recent HERV-K(HML-2) integrations and the hypothesized implication of some of the active loci in several diseases [13,167–169]. Groups belonging to class III were indeed generally less active. In comparison to Mommert *et al.* [111], who reported an abundant activation of class I and of all class III groups, HERV expression among classes showed some differences. These differences can be explained by i) the technologies and methods that have been used (*in vivo* vs *ex vivo,* microarray vs RNA-seq), and ii) the great differences in the basal transcriptional activity of each individual, which have been already observed in PBMCs [170]. Moreover, the use of the Vargiu *et al*. database [7] has allowed to analyze the expression of the most intact HERV proviruses, for which it is possible to hypothesize a higher likelihood of protein production. Of note, the pattern of counts distribution among classes is similar to the one obtained when considering the *hervgdb4* loci, and it is mostly in agreement with the information on HERV-H, HERV-K(HML-2), HERV-E and HERV-W transcriptional activity in PBMCs, reported in previous studies (37). However, such studies provided information on the overall expression of the above groups, but no information on the individual loci, here reported for the first time. Moreover, the expression of HERV-E has been previously reported to be characteristic of only a small percentage of the subjects analyzed, while a large portion of both *hervgdb4* loci and most intact HERVs was observed to be expressed in PBMCs in the present study.

Then, we studied the patterns of HERV/MaLR expression in immune response. We started with a model that mimicked a strong activation of the innate immune response, stimulated by high levels of the immunostimulant LPS. After that, we analyzed a context of adaptive immunity in response to the vaccine for hantaan virus, mimicking the following steps of a viral infection.

In the case of the LPS-stimulation, the analysis of the variability among the 30 samples analyzed showed differences between non-stimulated and LPS-stimulated samples. Indeed, in both PCA analysis and hierarchical clustering the not-stimulated and stimulated samples spread in two distinct clusters. However, a great interpersonal variability in the response of patients to LPS is also evident. This heterogeneity in the HERV and MaLR expression, as a consequence of LPS stimulus, is in line with the already observed strong inter-individual variability of gene expression in response to microbial agents [163,171].

Instead, the response to the vaccine was not sufficient to spread the samples in well-defined clusters, according to the variation of HERV expression after the vaccine administration, suggesting that HERV and MaLR patterns of expression are less involved in the adaptive immune response.

After both LPS stimulation and vaccine administration, we found HERV and MaLR modulated, and the majority of elements were up-regulated. Anyway, there were several differences between the two models, in terms of modulation. Indeed, the HERV and MaLR modulation was stronger after LPS stimulation than in response to vaccine administrations, and only a little portion of elements were differentially expressed in both the experimental models. This suggests that the modulation of specific HERV and MaLR loci may be different in different stages of the immune response.

Among the 15 most modulated *hervgdb4* fragments by LPS, 10 were co-localized with human genes, mostly related to innate immunity. None of the *hervgdb4* fragments that are most modulated after LPS are modulated after vaccine administrations. Moreover, we did not find *hervgdb4* fragments co-localized with genes involved in immunity above the most modulated after vaccines. Indeed, the HERV and MaLR modulation in innate immunity seems to be linked to genes induced by cytokines, and this induction lacks in vaccinated samples, according to the PCA of the 44 immune-related genes.

Among the 15 *hervgdb4* fragments most modulated by LPS, 10 are neighbor integrations of human genes that are also activated after the inflammatory response. Of particular interest is the identification of 3 solo LTRs localized outside the modulated genes, since the possible presence of promoters or polyadenylation signals may play a role in the regulation of the nearby gene. Specifically, the solo LTRs in chr2:151381023-151381962 and chr19:35887015-35887532 are integrated downstream of the TNF alpha-induced protein 6 gene (TNFAIP6) and NF-kappa-B inhibitor delta (NFKBID), respectively. Interestingly, while the most differentially expressed genes identified are mainly those coding for proteins that are positive regulators of immunity, such as IL1A and IL1B (data not shown), these data showed a strong up-regulation of LTRs-retrotransposon co-localized with genes coding for cellular inhibitors of these proteins. Hence, if on the one hand it has been suggested that a subset of HERVs hold TFBSs [172] that may increase their activation in immunity, on the other hand, the genes products of TNFAIP6 and NFKBID are potentially able to inhibit this phenomenon [173,174]. For this reason, such data underline the complexity of the relationship between the HERV/MaLR modulation and the immune response, especially if hypothesizing their active role in the regulation of the co-localized immune-related genes.

The results we obtained on the evaluation of the transcripts in the region of the higher concordant modulation of HERVs/MaLRs and the neighboring genes suggested that these elements are not included in chimeric transcripts with the genes. However, these data can not exclude a role of HERVs and MaLRs on the regulation of genes related to immunity, as more focused analyses would be necessary to investigate the phenomenon. Moreover, the strong positive linear correlation we observe between HERVs, MalRs and immune-related genes suggest that the expression of particular HERVs and MalRs could be used as markers of the immune activation and of the expression of immune-related genes.

Present data may give the basis to understand the HERV and MaLR modulation above the various stages of immunity. The analysis of the HERV and MaLR expression in innate immunity may help to understand the involvement in the regulation of immune functions, but further studies are needed to clarify these mechanisms. Similarly, further analyses are needed to characterize the HERV and MaLR expressed in adaptive immunity, to better understand their possible contribution to the immune response.

6. Conclusions

In a chapter of this thesis we characterized 66 HERV elements belonging to the HML-6 group, a member of the class II Beta-retrovirus-like, identifying two main HML-6 sequences clusters phylogenetically distinct, type 1 and 2, and an additional subdivision of type 1 in type 1a and 1b. Moreover, we found different sub-clusters of LTRs: LTR3 and LTR3A associated with only type 1a internal elements; LTR3B occurring with type 1b internal elements, and LTR3B_v occurring with type 2 members. We also observed structural differences among the two types of proviruses. Indeed, all type 2 elements showed the same deletion in *gag*, *pro* and *env*. While we did not study the possible mechanisms causing such deletions, further analyses should be performed to better understand their evolutionary pattern and, for example, if the elements derived by independent retroviral integrations or by duplication events.

We predicted, for the first time, a Rec domain within the *env* HML-6 squence, and we provided the first description of Rec in 23 HML-6 elements. However, it is still unclear whether HML-6 full-length Rec proteins are produced in any human tissue, and further analysis should be done to investigate possible involvements of this protein in human pathology and/or physiology.

We reported evidences about the context of insertion and co-localization of 19 HML-6 elements with functional human genes, including the sequence 16p11.2, whose 5'LTR overlapped the exon of one transcript variant of a cellular Zinc-finger up-regulated and involved in hepatocellular carcinoma.

In general, the present work provides the first complete overview of the HML-6 elements in GRCh37(hg19), describing the structure, phylogeny and genomic context of insertion of each locus. This information allows a better understanding of the genetics of one of the most interesting HERV groups in the human genome.

In the following chapter, we used RNA-seq transcriptome data from 15 healthy participants to a clinical trial injected with LPS, to identify expressed and modulated HERVs and MaLRs after activation of innate immune response. Then, we used RNA-seq transcriptome data from 19 subjects administered with an inactivated vaccine, to asses HERV and MaLR expression and modulation after activation of adaptive immune response.

Such RNA-seq based approach revealed the basal expression of HERVs and MaLRs in PBMCs. Moreover, results showed interpersonal differences in HERVs and MaLRs expression. We found that the HERV/MaLR expression patterns give a strong signature for the innate immune response, and a weaker signature for adaptive immune response. These results can be interesting for further studies that aim to identify specific HERVs and MaLRs possibly acting as biomarkers in immune-related diseases or immunocompromised conditions. For this reason, not only the HERV/MaLR transcripts need to be investigated, but it may be of interest to investigate also protein products.

We found 6,452 differentially expressed elements after LPS stimulation and 1,038 after the last vaccine administration, observing a general trend of up-regulation. Hence, we observed, for the first time, that HERV and MaLR loci are responsive to immune activation. Moreover, the HERVs and MaLRs modulated after LPS stimulation were not the same of those modulated after vaccine administration, suggesting specific patterns of activation in each stage of immunity.

Of note, the HERV/MaLR regulation after LPS stimulation was similar to that of co-localized and similarly modulated cellular genes. There is a strong correlation between HERV/MaLR expression and the expression of certain genes immune-related. This suggests possible applications of HERVs and MaLRs as biomarker in immunity settings, and it highlights the importance of investigate the role of LTR-retrotransposon expression in context of activation of the immune response and autoimmunity. Moreover, this co-localization is interesting for possible interactions between LTR-retrotransposons and the immune response. However, further analyses are required to evaluate if the HERV and MaLR modulation is an accidental and due to the modulation of neighbor genes, or if some of these elements are somehow involved in the regulation of the immune response.

Overall, these results allow to better assess the expression and modulation of LTR-retrotransposons expression in various stages of immunity, laying the bases for further studies that can clarify the impact and possible involvement of HERVs and MaLRs in the immune response.

Bibliografy

1. Baltimore, D. Expression of animal virus genomes. *Bacteriol. Rev.* **1971**, *35*, 235–241.

2. Lefkowitz, E.J.; Dempsey, D.M.; Hendrickson, R.C.; Orton, R.J.; Siddell, S.G.; Smith, D.B. Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* **2018**, *46*, D708–D717.

3. Family - Retroviridae. In *Virus Taxonomy*; King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J.B.T.-V.T., Eds.; Elsevier: San Diego, 2012; pp. 477–495 ISBN 978-0-12-384684-6.

4. Jern, P.; Coffin, J.M. Effects of Retroviruses on Host Genome Function. *Annu. Rev. Genet.* **2008**, *42*, 709–732.

5. Grandi, N.; Tramontano, E. HERV envelope proteins: Physiological role and pathogenic potential in cancer and autoimmunity. *Front. Microbiol.* **2018**, *9*, 1–26.

6. Smit, A.F.A. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res.* **1993**, *21*, 1863–1872.

7. Vargiu, L.; Rodriguez-Tomé, P.; Sperber, G.O.; Cadeddu, M.; Grandi, N.; Blikstad, V.; Tramontano, E.; Blomberg, J. Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology* **2016**, *13*, 2–29.

8. 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 Evol. Biol.* **2018**, *18*, 1–14.

9. Blomberg, J.; Benachenhou, F.; Blikstad, V.; Sperber, G.; Mayer, J. Classification and nomenclature of endogenous retroviral sequences (ERVs). Problems and recommendations. *Gene* **2009**, *448*, 115–123.

10. Becker, J.; Pérot, P.; Cheynet, V.; Oriol, G.; Mugnier, N.; Mommert, M.; Tabone, O.; Textoris, J.; Veyrieras, J.B.; Mallet, F. A comprehensive hybridization model allows whole HERV transcriptome profiling using high density microarray. *BMC Genomics* **2017**, *18*, 1–14.

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

12. Sperber, G.; Lövgren, A.; Eriksson, N.; 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**, *4*, 4–7.

13. Subramanian, R.P.; Wildschutte, J.H.; Russo, C.; Coffin, J.M. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* **2011**, *8*, 1–22.

14. 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.

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

16. Pisano, M.P.; Grandi, N.; Cadeddu, M.; Blomberg, J.; Tramontano, E. Comprehensive Characterization of the Human Endogenous Retrovirus HERV-K(HML-6) Group: Overview of Structure, Phylogeny, and Contribution to the Human Genome. *J. Virol.* **2019**, *93*, 1–19.

17. Grandi, N.; Cadeddu, M.; Pisano, M.P.; Esposito, F.; Blomberg, J.; Tramontano, E. Identification of a novel HERV-K(HML10): Comprehensive characterization and comparative analysis in non-human primates provide insights about HML10 proviruses structure and diffusion. *Mob. DNA* **2017**, *8*, 1–18.

18. 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.

19. Grandi, N.; Pisano, M.P.; Tramontano, E. The emerging field of human endogenous retroviruses: Understanding their physiological role and contribution to diseases. *Future Virol.* 2019, *14*, 441–444.

20. Cheynet, R.I.E.; Bouton, O.; Blond, J.; Lavillette, D. 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**, *4*, 3321–3329.

21. Sha, M.; Lee, X.; Li, X. ping; Veldman, G.M.; Finnerty, H.; Racie, L.; LaVallie, E.; Tang, X.Y.; Edouard, P.; Howes, S.; et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **2000**, *403*, 785–789.

22. Cegolon, L.; Salata, C.; Weiderpass, E.; Vineis, P.; Palù, G.; Mastrangelo, G. Human endogenous retroviruses and cancer prevention: Evidence and prospects. *BMC Cancer* **2013**, *13*, 4.

23. Kassiotis, G.; Stoye, J.P. Making a virtue of necessity: The pleiotropic role of human endogenous retroviruses in cancer. *Philos. Trans. R. Soc. B Biol. Sci.* **2017**, *372*.

24. 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*, 143–151.

25. Chen, T.; Meng, Z.; Gan, Y.; Wang, X.; Xu, F.; Gu, Y.; Xu, X.; Tang, J.; Zhou, H.; Zhang, X.; et al. The viral oncogene Np9 acts as a critical molecular switch for co-activating β-catenin, ERK, Akt and Notch1 and promoting the growth of human leukemia stem/progenitor cells. *Leukemia* **2013**, *27*, 1469–1478.

26. Gross, H.; Barth, S.; Pfuhl, T.; Willnecker, V.; Spurk, A.; Gurtsevitch, V.; Sauter, M.; Hu, B.; Noessner, E.; Mueller-Lantzsch, N.; et al. The NP9 protein encoded by the human endogenous retrovirus HERV-K(HML-2) negatively regulates gene activation of the Epstein-Barr virus nuclear antigen 2 (EBNA2). *Int. J. Cancer* **2011**, *129*, 1105–1115.

27. Hurst, T.P.; Magiorkinis, G. Activation of the innate immune response by endogenous retroviruses. *J. Gen. Virol.* **2015**, *96*, 1207–1218.

28. Grandi, N.; Tramontano, E. Human Endogenous Retroviruses Are Ancient Acquired Elements Still Shaping Innate Immune Responses. *Front. Immunol.* **2018**, *9*, 1–16.

29. 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.

30. Ruprecht, K.; Mayer, J. On the origin of a pathogenic HERV-W envelope protein present in multiple sclerosis lesions. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 19791–19792.

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

32. Grandi, N. Integrations and Their Mobilization by L1 Machinery : Contribution to the Human Transcriptome and Impact on the Host Physiopathology. *Viruses* **2017**, *9*, 1–37.

33. van de Lagemaat, L.N.; Medstrand, P.; Mager, D.L. Multiple effects govern endogenous retrovirus survival patterns in human gene introns. *Genome Biol.* **2006**, *7*, R86.1-R86.14.

34. Pi, W.; Zhu, X.; Wu, M.; Wang, Y.; Fulzele, S.; Eroglu, A.; Ling, J. Long-range function of an intergenic retrotransposon. *PNAs* **2010**, *107*, 12992–12997.

35. Samuelson, L.C.; Wiebauer, K.; Snow, C.M.; Meisler, M.H. Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. *Mol. Cell. Biol.* **1990**, *10*, 2513–2520.

36. Kamp, C. Two long homologous retroviral sequence blocks in proximal Yq11 cause AZFa microdeletions as a result of intrachromosomal recombination

events. *Hum. Mol. Genet.* **2000**, *9*, 2563–2572.

37. Chuma, S.; Pillai, R.S. Retrotransposon silencing by piRNAs: Ping-pong players mark their sub-cellular boundaries. *PLoS Genet.* **2009**, *5*, 1–3.

38. Armitage, A.E.; Katzourakis, A.; de Oliveira, T.; Welch, J.J.; Belshaw, R.; Bishop, K.N.; Kramer, B.; McMichael, A.J.; Rambaut, A.; Iversen, A.K.N. Conserved Footprints of APOBEC3G on Hypermutated Human Immunodeficiency Virus Type 1 and Human Endogenous Retrovirus HERV-K(HML2) Sequences. *J. Virol.* **2008**, *82*, 8743–8761.

39. Ito, J.; Gifford, R.J.; Sato, K. Retroviruses drive the rapid evolution of mammalian APOBEC3 genes. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 610–618.

40. Rowe, H.M.; Trono, D. Dynamic control of endogenous retroviruses during development. *Virology* **2011**, *411*, 273–287.

41. Hurst, T.P.; Magiorkinis, G. Epigenetic control of human endogenous retrovirus expression: Focus on regulation of long-terminal repeats (LTRs). *Viruses* 2017, *9*, 1–13.

42. Zhang, Y.; Li, T.; Preissl, S.; Amaral, M.L.; Grinstein, J.D.; Farah, E.N.; Destici, E.; Qiu, Y.; Hu, R.; Lee, A.Y.; et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat. Genet.* **2019**, *51*, 1380–1388 |.

43. Vincendeau, M.; Göttesdorfer, I.; Schreml, J.M.H.; Wetie, A.G.N.; Mayer, J.; Greenwood, A.D.; Helfer, M.; Kramer, S.; Seifarth, W.; Hadian, K.; et al. Modulation of human endogenous retrovirus (HERV) transcription during persistent and de novo HIV-1 infection. *Retrovirology* **2015**, *12*, 1–17.

44. Li, M.; Radvanyi, L.; Yin, B.; Li, J.; Chivukula, R.; Lin, K.; Lu, Y.; Shen, J.; Chang, D.Z.; Li, D.; et al. Down-regulation of human endogenous retrovirus type K (HERV- K) viral env RNA in pancreatic cancer cells decreases cell proliferation and tumor growth. *Clin. Cancer Res.* **2018**, *23*, 5892–5911.

45. Attig, J.; Young, G.R.; Stoye, J.P.; Kassiotis, G. Physiological and pathological transcriptional activation of endogenous retroelements assessed by RNA-sequencing of B lymphocytes. *Front. Microbiol.* **2017**, *8*, 1–11.

46. Tabone, O.; Mommert, M.; Jourdan, C.; Cerrato, E.; Legrand, M.; Lepape, A.; Allaouchiche, B.; Rimmelé, T.; Pachot, A.; Monneret, G.; et al. Endogenous retroviruses transcriptional modulation after severe infection, trauma and burn. *Front. Immunol.* **2019**, *10*, 1–12.

47. Turner, G.; Barbulescu, M.; Su, M.; Jensen-Seaman, M.I.; Kidd, K.K.; Lenz, J. Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.* **2001**, *11*, 1531–1535.

48. Wallace, A.D.; Wendt, G.A.; Barcellos, L.F.; de Smith, A.J.; Walsh, K.M.; Metayer, C.; Costello, J.F.; Wiemels, J.L.; Francis, S.S. To ERV is human: A phenotype-wide

scan linking polymorphic human endogenous retrovirus-K insertions to complex phenotypes. *Front. Genet.* **2018**, *9*, 1–14.

49. Lee, A.; Huntley, D.; Aiewsakun, P.; Kanda, R.K.; Lynn, C.; Tristem, M. Novel Denisovan and Neanderthal Retroviruses. *J. Virol.* **2014**, *88*, 12907–12909.

50. Agoni, L.; Golden, A.; Guha, C.; Lenz, J. Neandertal and Denisovan retroviruses. *Curr. Biol.* **2012**, *22*, R437–R438.

51. Marchi, E.; Kanapin, A.; Magiorkinis, G.; Belshaw, R. Unfixed Endogenous Retroviral Insertions in the Human Population. *J. Virol.* **2014**, *88*, 9529–9537.

52. Wildschutte, J.H.; Williams, Z.H.; Montesion, M.; Subramanian, R.P.; Kidd, J.M.; Coffin, J.M. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, E2326–E2334.

53. Li, W.; Id, L.L.; Malhotra, R.; Yang, L.; Acharya, R.; Id, M.P. A computational framework to assess genome-wide distribution of polymorphic human endogenous retrovirus-K In human populations. *PLoS Comput. Biol.* **2019**, *15*, 1–21.

54. Wallace, I.M.; Sullivan, O.O.; Higgins, D.G.; Notredame, C. M-Coffee : combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **2006**, *34*, 1692–1699.

55. Chang, T.C.; Goud, S.; Torcivia-Rodriguez, J.; Hu, Y.; Pan, Q.; Kahsay, R.; Blomberg, J.; Mazumder, R. Investigation of somatic single nucleotide variations in human endogenous retrovirus elements and their potential association with cancer. *PLoS One* **2019**, *14*, 1–23.

56. Montesion, M.; Williams, Z.H.; Subramanian, R.P.; Kuperwasser, C.; Coffin, J.M. Promoter expression of HERV-K (HML-2) provirus-derived sequences is related to LTR sequence variation and polymorphic transcription factor binding sites. *Retrovirology* **2018**, *15*, 1–16.

57. Faulkner, G.J.; Kimura, Y.; Daub, C.O.; Wani, S.; Plessy, C.; Irvine, K.M.; Schroder, K.; Cloonan, N.; Steptoe, A.L.; Lassmann, T.; et al. The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **2009**, *41*, 563–571.

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

59. Lu, X.; Sachs, F.; Ramsay, L.A.; Jacques, P.É.; Göke, J.; Bourque, G.; Ng, H.H. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat. Struct. Mol. Biol.* **2014**.

60. Grow, E.J.; Flynn, R.A.; Chavez, S.L.; Bayless, N.L.; Wesche, D.; Martin, L.; Ware, C.; Blish, C.A. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **2015**, *522*, 221–225.

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

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

63. Santoni, F.A.; Guerra, J.; Luban, J. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **2012**, *9*, 1.

64. Ruiz-Gonz??lez, I.; Xu, J.; Wang, X.; Burghardt, R.C.; Dunlap, K.A.; Bazer, F.W. Exosomes, endogenous retroviruses and toll-like receptors: Pregnancy recognition in ewes. *Reproduction* **2015**, *149*, 281–291.

65. Dodsworth, B.T.; Flynn, R.; Cowley, S.A. The current state of Naïve human pluripotency. *Stem Cells* **2015**, *33*, 3181–3186.

66. Lengronne, A.; Katou, Y.; Mori, S.; Yokobayashi, S.; Kelly, G.P.; Itoh, T.; Watanabe, Y.; Shirahige, K.; Uhlmann, F. Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature* **2004**, *430*, 573–578.

67. Raviram, R.; Rocha, P.P.; Luo, V.M.; Swanzey, E.; Miraldi, E.R.; Chuong, E.B.; Feschotte, C.; Bonneau, R.; Skok, J.A. Analysis of 3D genomic interactions identifies candidate host genes that transposable elements potentially regulate. *Genome Biol.* **2018**, *19*, 1–19.

68. Schmidt, D.; Schwalie, P.C.; Wilson, M.D.; Ballester, B.; Gonalves, Â.; Kutter, C.; Brown, G.D.; Marshall, A.; Flicek, P.; Odom, D.T. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **2012**, *148*, 335–348.

69. Friedli, M.; Trono, D. The Developmental Control of Transposable Elements and the Evolution of Higher Species. *Annu. Rev. Cell Dev. Biol.* **2015**, *31*, 429–451.

70. Leung, D.C.; Lorincz, M.C. Silencing of endogenous retroviruses: When and why do histone marks predominate? *Trends Biochem. Sci.* **2012**, *37*, 127–133.

71. Karimi, M.M.; Goyal, P.; Maksakova, I.A.; Bilenky, M.; Leung, D. DNA methylation and SETDB1 / H3K9me3 regulate predominantly distinct sets of genes , retroelements and chimaeric transcripts in mouse ES cells. *Cell Stem Cell* **2011**, *8*, 676–687.

72. Shi, H.; Strogantsev, R.; Takahashi, N.; Kazachenka, A.; Lorincz, M.C.; Hemberger, M.; Smith, A.C.F. ZFP57 regulation of transposable elements and gene expression within and beyond imprinted domains. *Epigenetics Chromatin* **2019**, *12*, 1–13.

73. Turelli, P.; Castro-Diaz, N.; Marzetta, F.; Kapopoulou, A.; Raclot, C.; Duc, J.; Tieng, V.; Quenneville, S.; Trono, D. Interplay of TRIM28 and DNA methylation in

controlling human endogenous retroelements. *Genome Res.* **2014**, *24*, 1260–1270.

74. Choi, Y.J.; Lin, C.; Risso, D.; Chen, S.; Tan, M.H.; Li, J.B.; Wu, Y.; Chen, C.; Xuan, Z.; Macfarlan, T.; et al. Deficiency of microRNA miR-34a expands cell fate potential in pluripotent stem cells. *Science (80-.).* **2017**, *355*, 1–26.

75. Ito, J.; Sugimoto, R.; Nakaoka, H.; Yamada, S.; Kimura, T.; Hayano, T.; Inoue, I. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLoS Genet.* **2017**, *13*, 1–33.

76. Haase, K.; Mösch, A.; Frishman, D. Differential expression analysis of human endogenous retroviruses based on ENCODE RNA-seq data. *BMC Med. Genomics* **2015**, *8*, 1–12.

77. Criscione, S.W.; Zhang, Y.; Thompson, W.; Sedivy, J.M.; Neretti, N. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **2014**, *15*, 1–17.

78. Id, T.N.; Autio, A.; Mishra, B.H.; Marttila, S. Aging-associated patterns in the expression of human endogenous retroviruses. *PLoS One* **2018**, *13*, 1–11.

79. Brocks, D.; Schmidt, C.R.; Daskalakis, M.; Jang, H.S.; Shah, N.M.; Li, D.; Li, J.; Zhang, B.; Hou, Y.; Laudato, S.; et al. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nat. Genet.* **2017**, *49*, 1052–1060.

80. Tobiasson, M.; Abdulkadir, H.; Lennartsson, A.; Marabita, F.; Paepe, A. De; Karimi, M.; Einarsdottir, E.; Grövdal, M.; Jansson, M.; Azenkoud, B.; et al. Comprehensive mapping of the effects of azacitidine on DNA methylation , repressive / permissive histone marks and gene expression in primary cells from patients with MDS and MDS- related disease. *Oncotarget* **2017**, *8*, 28812–28825.

81. Schmidt, N.; Domingues, P.; Golebiowski, F.; Patzina, C.; Tatham, M.H.; Hay, R.T. An influenza virus-triggered SUMO switch orchestrates co-opted endogenous retroviruses to stimulate host antiviral immunity. *PNAs* **2019**, *116*, 17399–17408.

82. Hunter, R.G.; Murakami, G.; Dewell, S.; Baker, M.E.R.; Datson, N.A. Acute stress and hippocampal histone H3 lysine 9 trimethylation , a retrotransposon silencing response. *PNAs* **2012**, *109*, 1–6.

83. Chuong, E.B.; Elde, N.C.; Feschotte, C.; Sanchez, A.; Trappier, S.G.; Mahy, B.W.; Peters, C.J.; Nichol, S.T.; Mohan, G.S.; Li, W.; et al. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science (80-.).* **2016**, *351*, 1083–1088.

84. White, C.H.; Beliakova-Bethell, N.; Lada, S.M.; Breen, M.S.; Hurst, T.P.; Spina, C.A.; Richman, D.D.; Frater, J.; Magiorkinis, G.; Woelk, C.H. Transcriptional modulation of human endogenous retroviruses in primary CD4+T cells

following vorinostat treatment. *Front. Immunol.* **2018**, *9*, 1–10.

85. Rajagopalan, D.; Tirado-Magallanes, R.; Bhatia, S.S.; Teo, W.S.; Sian, S.; Hora, S.; Lee, K.K.; Zhang, Y.; Jadhav, S.P.; Wu, Y.; et al. TIP60 represses activation of endogenous retroviral elements. *Nucleic Acids Res.* **2018**, *46*, 9456–9470.

86. Lock, F.E.; Babaian, A.; Zhang, Y.; Gagnier, L.; Kuah, S.; Weberling, A.; Karimi, M.M.; Mager, D.L. A novel isoform of IL-33 revealed by screening for transposable element promoted genes in human colorectal cancer. *PLoS One* **2017**, *12*, 1–30.

87. Johanning, G.L.; Malouf, G.G.; Zheng, X.; Esteva, F.J. Expression of human endogenous retrovirus-K is strongly associated with the basal-like breast cancer phenotype. *Sci. Rep.* **2017**, *7*, 1–11.

88. Siebenthall, K.T.; Miller, C.P.; Vierstra, J.D.; Mathieu, J.; Tretiakova, M.; Reynolds, A.; Sandstrom, R.; Rynes, E.; Haugen, E.; Johnson, A.; et al. Integrated epigenomic profiling reveals endogenous retrovirus reactivation in renal cell carcinoma. *EBioMedicine* **2019**, *41*, 427–442.

89. Id, F.L.; Sabunciyan, S.; Yolken, R.H.; Lee, D.; Kim, S. Transcription of human endogenous retroviruses in human brain by RNA-seq analysis. *PLoS One* **2019**, *14*, 1–13.

90. 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*, 13837–13852.

91. Reis, C.; Song, L.; Petri, M.; Sullivan, K.E. The SLE Transcriptome Exhibits Evidence of Chronic Endotoxin Exposure and Has Widespread Dysregulation of Non-Coding and Coding RNAs. *PLoS One* **2014**, *9*.

92. Tokuyama, M.; Kong, Y.; Song, E.; Jayewickreme, T.; Kang, I.; Iwasaki, A. ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses. *PNAs* **2018**, *115*, 12565–12572.

93. Karamitros, T.; Hurst, T.; Marchi, E.; Karamichali, E.; Georgopoulou, U.; Mentis, A.; Riepsaame, J.; Lin, A.; Paraskevis, D.; Hatzakis, A.; et al. Human endogenous retrovirus-K HML-2 integration within RASGRF2 is associated with intravenous drug abuse and modulates transcription in a cell-line model. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 10343–10439.

94. Mertz, J.A.; Simper, M.S.; Lozano, M.M.; Payne, S.M.; Dudley, J.P. Mouse Mammary Tumor Virus Encodes a Self-Regulatory RNA Export Protein and Is a Complex Retrovirus. *J. Virol.* **2005**, *79*, 14737–14747.

95. Magin, C.; Löwer, R.; Löwer, J. cORF and RcRE, the Rev/Rex and RRE/RxRE homologues of the human endogenous retrovirus family HTDV/HERV-K. *J. Virol.* **1999**, *73*, 9496–507.

96. Mayer, J.; Ehlhardt, S.; Seifert, M.; Sauter, M.; Müller-Lantzsch, N.; Mehraein, Y.; Zang, K.D.; Meese, E. Human endogenous retrovirus HERV-K(HML-2) proviruses with Rec protein coding capacity and transcriptional activity. *Virology* **2004**, *322*, 190–198.

97. Armbruester, V.; Sauter, M.; Roemer, K.; Best, B.; Hahn, S.; Nty, A.; Schmid, A.; Philipp, S.; Mueller, A.; Mueller-lantzsch, N. Np9 Protein of Human Endogenous Retrovirus K Interacts with Ligand of Numb Protein X Np9 Protein of Human Endogenous Retrovirus K Interacts with Ligand of Numb Protein X. *J. Virol.* **2004**, *78*, 10310–10319.

98. Medstrand, P.; Mager, D.L.; Yin, H.; Dietrich, U.; Blomberg, J. Structure and genomic organization of a novel human endogenous retrovirus family: HERV-K (HML-6). *J. Gen. Virol.* **1997**, *78*, 1731–1744.

99. Yin, H.; Medstrand, P.; Kristofferson, A.; Dietrich, U.; Åman, P.; Blomberg, J. Characterization of Human MMTV-like (HML) Elements Similar to a Sequence That Was Highly Expressed in a Human Breast Cancer: Further Definition of the HML-6 Group. *Virology* **1999**, *256*, 22–35.

100. Mayer, J.; Meese, E.U. Presence of dUTPase in the Various Human Endogenous Retrovirus K (HERV-K) Families. *J. Mol. Evol.* **2003**, *57*, 642–649.

101. 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.

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

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

104. Chiappinelli, K.B.; Strissel, P.L.; Desrichard, A.; Li, H.; Henke, C.; Akman, B.; Hein, A.; Rote, N.S.; Cope, L.M.; Snyder, A.; et al. Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* **2015**, *162*, 974–986.

105. Aswad, A.; Katzourakis, A. Paleovirology and virally derived immunity. *Trends Ecol. Evol.* **2012**, *27*, 627–636.

106. 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.

107. Meylan, F.; De Smedt, M.; Leclercq, G.; Plum, J.; Leupin, O.; Marguerat, S.; Conrad, B. Negative thymocyte selection to HERV-K18 superantigens in humans. *Blood*

2005, *105*, 4377–4382.

108. Hummel, J.; Kämmerer, U.; Müller, N.; Avota, E.; Schneider-Schaulies, S. Human endogenous retrovirus envelope proteins target dendritic cells to suppress T-cell activation. *Eur. J. Immunol.* **2015**, *45*, 1748–1759.

109. Madeira, A.; Burgelin, I.; Perron, H.; Curtin, F.; Lang, A.B.; 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.

110. De La Hera, B.; Varadé, J.; García-Montojo, M.; Alcina, A.; Fedetz, M.; Alloza, I.; Astobiza, I.; Leyva, L.; Fernández, O.; Izquierdo, G.; et al. Human endogenous retrovirus HERV-Fc1 association with multiple sclerosis susceptibility: A meta-analysis. *PLoS One* **2014**, *9*, 1–6.

111. Mommert, M.; Tabone, O.; Oriol, G.; Cerrato, E.; Guichard, A.; Naville, M.; Fournier, P.; Volff, J.N.; Pachot, A.; Monneret, G.; et al. LTR-retrotransposon transcriptome modulation in response to endotoxin-induced stress in PBMCs. *BMC Genomics* **2018**, *19*, 1–17.

112. Johnston, J.B.; Silva, C.; Holden, J.; Warren, K.G.; Clark, A.W.; Power, C. Monocyte activation and differentiation augment human endogenous retrovirus expression: Implications for inflammatory brain diseases. *Ann. Neurol.* **2001**, *50*, 434–442.

113. Voisset, C.; Weiss, R.A.; Griffiths, D.J. Human RNA “Rumor” Viruses: the Search for Novel Human Retroviruses in Chronic Disease. *Microbiol. Mol. Biol. Rev.* **2008**, *72*, 157–196.

114. Magiorkinis, G.; Belshaw, R.; Katzourakis, A. ‘There and back again’: revisiting the pathophysiological roles of human endogenous retroviruses in the post-genomic era. *Philos. R. Soc. B* **2013**, *368*, 1–12.

115. Reuter, J.A.; Spacek, D.; Snyder, M.P. High-Throughput Sequencing Technologies. *Mol Cell.* **2015**, *58*, 586–597.

116. Churko, J.M.; Mantalas, G.L.; Snyder, M.P.; Wu, J.C. Overview of High Throughput sequencing. *Clin. Res.* **2013**, *112*, 1–26.

117. Sudmant, P.H.; Rausch, T.; Gardner, E.J.; Handsaker, R.E.; Abyzov, A.; Huddleston, J.; Zhang, Y.; Ye, K.; Jun, G.; Fritz, M.H.Y.; et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **2015**, *526*, 75–81.

118. Auton, A.; Abecasis, G.R.; Altshuler, D.M.; Durbin, R.M.; Bentley, D.R.; Chakravarti, A.; Clark, A.G.; Donnelly, P.; Eichler, E.E.; Flicek, P.; et al. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74.

119. Sanchez-Vega, F.; Mina, M.; Armenia, J.; Schultz, N. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **2018**, *173*, 321–337.

120. Campbell, P.J.; Getz, G.; Korbel, J.O.; Stuart, J.M.; Jennings, J.L.; Stein, L.D.; Perry,

M.D.; Nahal-Bose, H.K.; Ouellette, B.F.F.; Li, C.C.H.; et al. Pan-cancer analysis of whole genomes. *Nature* **2020**, *578*, 82–93.

121. The International Cancer Genome Consortium International network of cancer genome projects. *Nature* **2010**, *464*, 993–998.

122. Chen, X.; Li, D.; Birol, I. ERVcaller: Identifying polymorphic endogenous retrovirus and other transposable element insertions using whole-genome sequencing data. *Bioinformatics* **2019**, *35*, 3913–3922.

123. Sun, W.; Hu, Y. eQTL Mapping Using RNA-seq Data. *Stat. Biosci.* **2013**, *5*, 198–219.

124. Lieberman-Aiden, E.; van Berkum, N.L.; Dekker, J. Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science (80-.).* **2009**, *326*, 289–293.

125. Sokol, M.; Jessen, K.M.; Pedersen, F.S. Utility of next-generation RNA-sequencing in identifying chimeric transcription involving human endogenous retroviruses. *Apmis* **2016**, *124*, 127–139.

126. Sokol, M.; Jessen, K.M.; Pedersen, F.S. Human endogenous retroviruses sustain complex and cooperative regulation of gene-containing loci and unannotated megabase-sized regions. *Retrovirology* **2015**, *12*, 1–11.

127. Gosenca, D.; Gabriel, U.; Steidler, A.; Mayer, J.; Diem, O.; Erben, P.; Hofmann, W.; Seifarth, W.; Fabarius, A.; Leib-mo, C. HERV-E-Mediated Modulation of PLA2G4A Transcription in Urothelial Carcinoma. *PLoS One* **2012**, *7*, 1–15.

128. Carding, S.R.; Hoyles, N.D.L. Review article : the human intestinal virome in health and disease. *Aliment. Pharmacol. Ther.* **2017**, *46*, 800–815.

129. Karolchik, D.; Barber, G.P.; Casper, J.; Clawson, H.; Cline, M.S.; Diekhans, M.; Dreszer, T.R.; Fujita, P.A.; Guruvadoo, L.; Haeussler, M.; et al. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **2014**, *42*, 764–770.

130. Kent, W.J.; Sugnet, C.W.; Furey, T.S.; Roskin, K.M. The Human Genome Browser at UCSC W. *J. Med. Chem.* **1976**, *19*, 1228–31.

131. Hubley, R.; Finn, R.D.; Clements, J.; Eddy, S.R.; Jones, T.A.; Bao, W.; Smit, A.F.A.; Wheeler, T.J. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **2016**, *44*, D81–D89.

132. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780.

133. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C.; et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012**, *28*, 1647–1649.

134. Darriba, D.; Taboada, G.L.; Doallo, R.; Posada, D. Europe PMC Funders Group

jModelTest 2 : more models , new heuristics and high- performance computing. *Nat Methods* **2015**, *9*, 6–9.

135. 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.

136. Guindon S., G.O. A Simple , Fast , and Accurate Algorithm to Estimate Large Phylogenies. *Syst. Biol.* **2003**, *52*, 696–704.

137. Marchler-Bauer, A.; Bo, Y.; Han, L.; He, J.; Lanczycki, C.J.; Lu, S.; Chitsaz, F.; Derbyshire, M.K.; Geer, R.C.; Gonzales, N.R.; et al. CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **2017**, *45*, D200–D203.

138. Barnett, D.W.; Garrison, E.K.; Quinlan, A.R.; Strömberg, M.P.; Marth, G.T.; Api, T. BamTools : a C ++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **2011**, *27*, 1691–1692.

139. Anders, S.; Pyl, P.T.; Huber, W. Genome analysis HTSeq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* **2015**, *31*, 166–169.

140. Harrow, J.; Frankish, A.; Gonzalez, J.M.; Tapanari, E.; Diekhans, M.; Kokocinski, F. GENCODE: The Reference Human Genome Annotation for The ENCODE Project. *Genome Res* **2012**, *22*, 1760–1774.

141. Aboyoun, P.; Carlson, M.; Lawrence, M.; Huber, W.; Gentleman, R.; Morgan, M.T.; Carey, V.J. Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **2013**, *9*, 1–10.

142. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 1–21.

143. Hochberg, Y.B. and Y. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing Author (s): Yoav Benjamini and Yosef Hochberg Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 57 , No . 1 Published by : *J. R. Stat. Soc.* **1995**, *57*, 289–300.

144. Aken, B.L.; Ayling, S.; Barrell, D.; Clarke, L.; Curwen, V.; Fairley, S.; Fernandez Banet, J.; Billis, K.; García Girón, C.; Hourlier, T.; et al. The Ensembl gene annotation system. *Database (Oxford).* **2016**, *2016*, 1–19.

145. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652.

146. Wu, T.D.; Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **2005**, *21*, 1859–1875.

147. Robinson, J.T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative genomics viewer. *Nat. Biotechnol.* **2011**, *29*, 24–26.

148. Tristem, M. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *J. Virol.* **2000**, *74*, 3715–30.

149. Medstrand, P.; Blomberg, J. Characterization of novel reverse transcriptase encoding human endogenous retroviral sequences similar to type A and type B retroviruses: differential transcription in normal human tissues. *J. Virol.* **1993**, *67*, 6778–6787.

150. Lebedev, Y.B.; Belonovitch, O.S.; Zybrova, N. V; Khil, P.P.; Kurdyukov, S.G.; Vinogradova, T. V; Hunsmann, G.; Sverdlov, E.D. Di ff erences in HERV-K LTR insertions in orthologous loci of humans and great apes. **2000**, *247*, 265–277.

151. Johnson, W.E.; Coffin, J.M. Constructing primate phylogenies from ancient retrovirus sequences. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 10254–60.

152. Doxiadis, G.G.M.; de Groot, N.; Bontrop, R.E. Impact of Endogenous Intronic Retroviruses on Major Histocompatibility Complex Class II Diversity and Stability. *J. Virol.* **2008**, *82*, 6667–6677.

153. Mack, M.; Bender, K.; Schneider, P.M. Detection of retroviral antisense transcripts and promoter activity of the HERV-K (C4) insertion in the MHC class III region. *Immunogenetics* **2004**, *56*, 321–332.

154. Shigematsu, S.; Fukuda, S.; Nakayama, H.; Inoue, H.; Hiasa, Y.; Onji, M.; Higashiyama, S. ZNF689 suppresses apoptosis of hepatocellular carcinoma cells through the down-regulation of Bcl-2 family members. *Exp. Cell Res.* **2011**, *317*, 1851–1859.

155. Jern, P.; Sperber, G.O.; Blomberg, J. Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. *Retrovirology* **2005**, *2*, 1–12.

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

157. Boese, A.; Sauter, M.; Galli, U.; Best, B.; Herbst, H.; Mayer, J.; Kremmer, E.; Roemer, K.; Mueller-Lantzsch, N. Human endogenous retrovirus protein cORF supports cell transformation and associates with the promyelocytic leukemia zinc finger protein. *Oncogene* **2000**, *19*, 4328–4336.

158. Armbruester, V.; Sauter, M.; Krautkraemer, E.; Meese, E.; Kleiman, A.; Best, B.; Roemer, K.; Mueller-lantzsch, N. A Novel Gene from the Human Endogenous Retrovirus K Expressed in Transformed Cells 1. **2002**, 1800–1807.

159. Carithers, L.J.; Ardlie, K.; Barcus, M.; Branton, P.A.; Britton, A.; Buia, S.A.; Compton, C.C.; DeLuca, D.S.; Peter-Demchok, J.; Gelfand, E.T.; et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv. Biobank.* **2015**, *13*, 311–319.

160. Sjøttem, E.V.A.; Anderssen, S.; Johansen, T. The Promoter Activity of Long Terminal Repeats of the HERV-H Family of Human Retrovirus-Like Elements Is Critically Dependent on Sp1 Family Proteins Interacting with a GC / GT Box Located Immediately 3 J to the TATA Box. *J. Virol.* **1996**, *70*, 188–198.

161. Benachenhou, F.; Jern, P.; Oja, M.; Sperber, G.; Blikstad, V.; Somervuo, P.; Kaski, S.; Blomberg, J. Evolutionary conservation of orthoretroviral long terminal repeats (LTRs) and ab initio detection of single LTRs in genomic data. *PLoS One* **2009**, *4*, e5179.

162. Waugh, C.A.; Hanger, J.; Loader, J.; King, A.; Hobbs, M.; Johnson, R.; Timms, P. Infection with koala retrovirus subgroup B (KoRV-B), but not KoRV-A, is associated with chlamydial disease in free-ranging koalas (Phascolarctos cinereus). *Sci. Rep.* **2017**, *7*, 1–11.

163. Urrutia, A.; Duffy, D.; Rouilly, V.; Posseme, C.; Djebali, R.; Illanes, G.; Libri, V.; Albaud, B.; Gentien, D.; Piasecka, B.; et al. Standardized Whole-Blood Transcriptional Profiling Enables the Deconvolution of Complex Induced Immune Responses. *Cell Rep.* **2016**, *16*, 2777–2791.

164. Arend, W.P.; Gabay, C. Physiologic role of interleukin-1 receptor antagonist. *Arthritis Res.* **2000**, *2*, 2–5.

165. Lee, T.H.; Wisniewski, H.; Vildek, J. A Novel Secretory Tumor Necrosis Factor-inducible Protein (TSG-6) Is a Member of the Family of Hyaluronate Binding Proteins , Closely Related to the Adhesion Receptor CD44. *J. Cell Biol. J. Cell Biol.* **1992**, *116*, 545–557.

166. Khan, A.; Shin, O.S.; Na, J.; Kim, J.K.; Seong, R.K.; Park, M.S.; Noh, J.Y.; Song, J.Y.; Cheong, H.J.; Park, Y.H.; et al. A Systems Vaccinology Approach Reveals the Mechanisms of Immunogenic Responses to Hantavax Vaccination in Humans. *Sci. Rep.* **2019**, *9*, 1–14.

167. George R. Young, Sandra N. Terry, Lara Manganaro, Alvaro Cuesta-Dominguez, G.D.; Dabeiba, B.-R.; Laura Campisi, Ana Fernandez-Sesma, R.S.; Viviana Simon, L.C.F.M. HIV-1 Infection of Primary CD4+ T Cells Regulates the Expression of Specific Human Endogenous Retrovirus HERV-K (HML-2) Elements. *J. Virol.* **2018**, *92*, 1–13.

168. Mayer, J.; Harz, C.; Sanchez, L.; Pereira, G.C.; Maldener, E.; Heras, S.R.; Ostrow, L.W.; Ravits, J.; Batra, R.; Meese, E.; et al. Transcriptional profiling of HERV-K (HML-2) in amyotrophic lateral sclerosis and potential implications for expression of HML-2 proteins. *Mol. Neurodegener.* **2018**, 1–25.

169. Ma, W.; Hong, Z.; Liu, H.; Chen, X.; Ding, L.; Liu, Z.; Zhou, F.; Yuan, Y. Human

Endogenous retroviruses-k (HML-2) expression is correlated with prognosis and progress of hepatocellular carcinoma. *Biomed Res. Int.* **2016**, *2016*.

170. 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*, 1–9.

171. Ferguson, J.F.; Patel, P.N.; Shah, R.Y.; Mulvey, C.K.; Gadi, R.; Nijjar, P.S.; Usman, H.M.; Mehta, N.N.; Shah, R.; Master, S.R.; et al. Race and gender variation in response to evoked inflammation. *J. Transl. Med. 2013,* **2013**, *11*, 1–9.

172. Chuong, E.B.; Elde, N.C.; Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science (80-.).* **2016**, *351*, 1083–1087.

173. Dyer, D.P.; Salanga, C.L.; Johns, S.C.; Valdambrini, E.; Fuster, M.M.; Milner, C.M.; Day, A.J.; Handel, T.M. The Anti-inflammatory Protein TSG-6 Regulates Chemokine Function by Inhibiting Chemokine / Glycosaminoglycan. *J. Biol. Chem.* **2016**, *291*, 12627–12640.

174. Tao, Z.; Fusco, A.; Huang, D.; Gupta, K.; Young, D.; Ware, C.F. p100 / I κ B δ sequesters and inhibits NF- κ B through kappaBsome formation. *PNAS* **2014**, *111*, 15946–15951.

Acknowledgments

I would like to thank all the people that have been involved in the various studies included in this thesis. First of all, my supervisor, Enzo Tramontano, who has been a patient, attentive and excellent guide over the last three years. Professor Jonas Blomberg, who is sadly passed away too soon. Professors François Mallet and Julien Textoris for their supervision and hospitality over the time I spent in Lyon. Many François Mallet's suggestions have been essential for a large part of this thesis. I Also would like to thank Nicole Grandi, colleague and friend, precious collaborator and supporter. Olivier Tabone and Maxime Bodinier, for all the help and training on the bioinformatics applications. Many thanks to Prof. Giacomo Diaz for valuable critical revision of the statistic of this thesis.

Special thanks to all the people from the Laboratory of Molecular Virology in Cagliari and from the HEH in Lyon, in particular Francesca, Angela, Nicole, Elisa, Alessia, Annalaura, Dina and Asma. Their company have made special (and fun) these three years.

I dedicate this thesis to all the people in my life, and especially to my family. To my parents, which always support me, and which made me who I am now. To my sister, so distant and yet so close to my heart. To my brother, whose creativity is a source of inspiration. Finally, to Federico, who is always present and lovely, simply making easier and more beauty everything.

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (CCI 2014IT16M2OP005),
Fondo Sociale Europeo, Azione I.1 “Dottorati Innovativi con caratterizzazione Industriale”

UNIONE EUROPEA
Fondo Sociale Europeo

Ministero dell' Istruzione,
dell'Università e della Ricerca

PON
RICERCA
E INNOVAZIONE
2014 - 2020