Comprehensive identification and characterization of HERV-K (HML-9) group in the human genome

Lin Li (dearwood@sina.com)
Beijing Institute of Microbiology and Epidemiology
Lei Jia
Beijing Institute of Microbiology and Epidemiology
Mengying Liu
Beijing University of Chemical Technology
Hanping Li
Beijing Institute of Microbiology and Epidemiology
Yongjian Liu
Beijing Institute of Microbiology and Epidemiology
Jingwan Han
Beijing Institute of Microbiology and Epidemiology
Xiuli Zhai
Beijing Institute of Microbiology and Epidemiology
Xiaolin Wang
Beijing Institute of Microbiology and Epidemiology
Tianyi Li
Beijing Institute of Microbiology and Epidemiology
Jingyun Li
Beijing Institute of Microbiology and Epidemiology
Bohan Zhang
Beijing Institute of Microbiology and Epidemiology
Changyuan Yu
Beijing University of Chemical Technology

Research Article

Keywords: Human endogenous retrovirus, HML-9, BLAT, GRCh38/hg38, gene regulation

Posted Date: March 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1426473/v1

Abstract

Background: Human endogenous retroviruses (HERVs) result from ancestral infections by exogenous retroviruses that became incorporated into germ-line DNA and evolutionary fixed in the human genome. HERVs could vertically transmit in a Mendelian fashion and stable maintenance in the human genome which are estimated to comprise about 8%. HERV-K (HML1-10) transcription has been confirmed to be associated with a variety of diseases, such as breast cancer, lung cancer, prostate cancer, melanoma, rheumatoid arthritis, and amyotrophic lateral sclerosis. However, the poorly characterization of HML-9 hinders a detailed understanding of the expression regulation of this family in human health and its actual impact on host genomes. In the light of this, the definition of a precise and updated HERV-K HML-9 genomic map is urgently needed.

Results: We report a comprehensive analysis of HERV-K HML-9 sequences presence and distribution within the human genome, with a detailed description of the different structural and phylogenetic aspects characterizing the group. A total of 23 proviruses and 47 solo LTR elements were characterized with a detailed description of provirus structure, integration time, potentially regulated genes, transcription factor binding sites, and primer binding site features. The integration time results showed that the HML-9 elements found in the human genome have been integrated into the primate lineage between 37.5 and 151.5 million years ago (mya).

Conclusion: The results have finally clarified the composition of HML-9, providing an exhaustive background for subsequent functional studies.

Background

The human genome contains many genetic loci evolved as a result of infection by different types of retroviruses, and about 45% of the human genome is composed of or derived from virus-like transposable elements (TEs) [1–3]. The most abundant TEs in the human genome are reverse transcription factors (REs) amplified by a copy-paste mechanism [3, 4]. One class of REs, human endogenous retroviruses (HERVs), results of ancestral infections by exogenous retroviruses that became incorporated into germ-line DNA and evolutionary fixed in the genome, which are estimated to comprise about 8% [5]. HERVs can be transmitted vertically as proviruses in a Mendelian fashion but are not inherently infectious [6–8]. HERVs are structurally similar to proviruses of common retroviruses where the *gag*, *pol*, and *env* genes are flanked by two long terminal repeats (LTRs) that act as promoters [5]. Most HERV families have pro genes, but some families have none, like HERV-K HML10 [9]. Even most *gag*, *pro*, *pol*, and *env* remain, they are usually inactive due to the accumulation of substitutions, deletions, and insertions.

Human endogenous retroviruses (HERVs) have been divided into three classes, of which Class I consists of Gamma retrovirus-like, Class II of Beta retrovirus-like, and Class III of vaguely Spuma retrovirus-like elements [10]. The classification of HERVs is complex, with different classification systems coexisting. In addition to the system based on identities in the *pol* sequence, another system exists based on the tRNA molecule used by retroviruses as a primer in replication. The primer binding site (PBS) regions of Class II are complementary to lysine (K) tRNA molecules which have been designated as HERV-K [11]. Among a variety of human endogenous retroviral families, the HERV-K was the latest acquired by the human species [12]. HERV-K is the most complete and biologically active family in the human genome and is closely related to many cancers and neurological diseases [6, 13–15]. HERV-K is divided into subfamilies from HML-1 through HML-10. These proviruses appeared about 30 ~ 35 million years ago [16, 17].

HERVs are usually enriched outside transcription units. The minority of HERV elements that are present within transcription units show a strong orientation bias. The sense-oriented polyadenylation signal in the LTR causes premature transcript termination, leading to the integration of LTRs in introns which are preferentially oriented antisense to the enclosing gene [10, 18–23]. Compared to other regions of the human genome, the male-specific regions of the Y chromosome (MSY) accumulate higher densities of HERVs and the associated sequences [24]. The integrated LTR elements have been shown to influence gene regulation by providing regulatory elements such as enhancers, promoters, splice- and polyadenylation sites for various host genes [4].

The first study of the relationship between reverse transcriptase (RT) protein of HERV expression and cancer was reported in the early 1970s [25, 26]. Currently, HERV-K transcription has been confirmed to be associated with a variety of diseases, such as breast cancer, lung cancer, prostate cancer, melanoma, rheumatoid arthritis, and amyotrophic lateral sclerosis [27–34]. The *env* gene of HERV-K can produce two oncoproteins derived through alternative *env* mRNA splicing, called Np9 and Rec. Both oncoproteins can induce carcinogenesis by the dysregulation of essential cellular pathways, leading to the inhibition of cellular growth, proliferation, and apoptosis [35, 36]. Because HERV protein is expressed in a variety of cancers, it belongs to tumor-specific antigens. They can impact both innate and adaptive immune responses, leading to B- and T-cell stimulation and activation, inducing antibodies and cytotoxic T-cell responses. In vitro and animal models, antibodies to HERV-K showed a role in inhibiting tumor growth [37]. HERVs are transcriptionally repressed in adult tissues through DNMT1-dependent cytosine methylation, which contributed to blocking the expression of its nucleic acids and proteins and potentially trigger an autoimmune response. The reactivations of HERVs in cells are also associated with autoimmune diseases such as multiple sclerosis [38, 39].

Characterization of the genomic distribution of the HML-9 group is critical to understanding its portrait and its relationship with diseases. Currently, there are very limited characterizations and research for HML-9 at present. In the light of this, the definition of a precise and updated HML-9 genomic map is a pressing need to better evaluate their role in human health.

Here we reported a comprehensive analysis of HERV-K HML-9 elements' presence and distribution within the human genome, with a detailed identification of different structural and phylogenetic aspects characterizing the group. Additionally, we analyzed provirus integration time and potentially regulated genes. Overall, the results have finally clarified the characterization of HML-9 and provided an exhaustive background for subsequent functional studies.

Materials And Methods

1. HML-9 identification, localization, and genomic distribution

To identify the HML-9 provirus and solo LTR distribution into the human genome, we performed the HML-9 identification by using the Genome Reference Consortium released Dec. 2013 assembly GRCh38/hg38 as the human genome reference. A traditional BLAT search tool [40] in the UCSC Genome Browser database [41] was used to identify the integrated HML-9 elements. DNA BLAT works by keeping an index of the entire genome in memory. The index consists of all overlapping 11-mers stepping by 5 except for those heavily involved in repeats (http://genome.ucsc.edu/cgi-bin/hgBlat). The assembled LTR14C-HERVK14C-LTR14C sequence was used as a query. Additionally, the expected distribution of the HML-9 loci in each chromosome were predicted according to the formula: e = Cl ∗ n / Tl (e is the number of integrations expected in the chromosome, Cl represents the ungapped length of the chromosome, n is the total number of actual HML-9 loci identified in the human genome, and Tl represents the sum ungapped length of all chromosome) [42]. The variation of the expected integrations was compared to the actual number of HML-9 loci through a chi-square (χ^2) test, and its statistical significance was estimated through the *p*-value.

2. Structural Characterization

The identified HML-9 elements were aligned to the proviral reference LTR14C-HERVK14C-LTR14C. Alignments have been analyzed on the BioEdit software platform [43]. All insertions and deletions were annotated.

3. Phylogenetic analyses

Maximum likelihood (ML) phylogenetic trees were built with Mega 7 [44] to confirm the assignment of the identified HML-9 elements. The 44 out of 47 solo LTRs sequences longer than 90% of the LTR14C and 5 out of 23 provirus sequences longer than 80% of LTR14C-HERVK14C-LTR14C were used to construct phylogenetic trees. The best-fitting models of nucleotide substitution for solo LTRs and full-length proviruses were calculated as K2 + G and GTR + G + I by Model Selection function in Mega7, respectively. For the 4 coding regions, elements longer than 90% of the corresponding section of HML-9 were screened to construct phylogenetic trees. According to the model selection function of Mega 7, the best-fitting models of nucleotide substitution for *gag*, *pro*, *pol*, and *env* analysis were HKY + G + I, GTR + G + I, GTR + G, and HKY + G, respectively. Tree topologies were searched using the nearest neighbor interchange (NNI) procedure. The confidence of each node in phylogenetic trees was determined using the bootstrap test with 500 bootstrap replicates. The final ML trees were visualized using iTOL [45].

4. Estimation of the integration of HML-9

To estimate the time of integration, we used the substitution rate (0.2%/nucleotides/million year) to assess the action of divergence on each HML-9 element [46]. The D is the percentage of divergent nucleotides and the D of each HML-9 member was estimated between (1) 5' and 3' LTRs of each provirus, (2) each HML-9 internal element *gag*, *pro*, *pol*, and *env* genes and its generated consensus. The divergence values were estimated with MEGA7. For 4 internal regions, the integration time was calculated based on the formula T = D/0.2, in which T represents the estimated time of integration (in million years). For the flanking LTR regions, the provirus integration time was calculated based on the formula T = D/0.2/2.

5. Function prediction of cis-regulatory regions and enrichment analysis

Non-coding regions typically lack biological functions annotation. To understand the biological significance of both HML-9 solo LTRs and provirus LTRs, an annotated analysis of adjacent genes of LTRs based on Genomic Regions Enrichment of Annotations Tool (GREAT) was performed [47]. The association rule is as follows: Basal + extension: 5000 bp upstream, 1000 bp downstream, 1000000 bp max extension, curated regulatory domains included. After screening out the potential regulatory genes, the WEB-based Gene SeT AnaLysis Toolkit (WebGestalt) [48] was used to analyze the functional enrichment (http://www.webgestalt.org), which is crucial in interpreting the list of interesting genes. The enrichment method used in the current work is ORA. Parameters for the enrichment analysis: Minimum number of IDs in the category: 5. Maximum number of IDs in the category: 2000. FDR Method: BH. Significance Level: Top 10.

6. In silico examination of the conserved transcription factor binding sites

The transcriptional binding sites of the HML-9 LTR consensus reference sequence were predicted on the JASPAR (https://jaspar.genereg.net/) database. The taxon was vertebrates, the species was homo sapiens. The ChIP-seq data were selected to predict transcription factors with a relative profile score threshold ≥ 95%. The alignment and annotation of the HML-9 LTR reference sequence with 4 proviral sequences (the length of 5'LTR > 90%) were performed using Geneious software [49].

7. Primer binding site feature representation

The composition of the PBS of 5 near-full-length proviruses (length > 80%) and the HML-9 reference sequence were all analyzed using Mega 7 and BioEdit. The grade of conservation at each position was represented with a logo built from WebLogo at http://weblogo.berkeley.edu [50].

Results

1. HML-9 elements identification, localization, and actual distribution in hg38

According to the BLAT results of LTR14C-HERVK14C-LTR14C in GRCh38/hg38, we characterized a total of 23 HERV-K HML-9 provirus elements. Each HML-9 element is named according to the genomic locus of integration (Table 1). It has been identified that the average length of the provirus is 5473 bp. Among them, 6 elements are longer than 70% of the full length of the reference, 9 elements are between 40–70% in length, and the remaining 8 sequences are between 17.11% and 34.26% of the reference sequence in length. For solo LTRs, 47 solo LTR elements of HERV-K HML-9 were characterized in total. Of these, 44 solo LTRs (93.62%) are longer than 90% of the representative reference LTR14C in length. The nucleotide sequence of each element was shown in Supplementary dataset 1–2. The whole HML-9 element distribution was displayed based on Ensembl (www.ensembl.org) (Fig. 1A).

Table 1
HML-9 provirus distribution.

Number	Locus	Chromosome	Strand	Position start	Position end	Length (bp)	Match + mismatch(bp)/full length(bp)	Range	Qgap(bp)/match + mismatch + Qgap(bp)	Insertic or deletio
1.	16p12.3	chr16	-	19393581	19402152	8572	96.00%	(90%-100%)	1.01%	NA
2.	2p12	chr2	+	82022660	82031279	8620	95.91%	(90%-100%)	1.13%	NA
3.	15q21.1	chr15	-	45234477	45243073	8597	95.34%	(90%-100%)	1.85%	NA
4.	8p11.1	chr8	-	43694016	43702583	8568	95.10%	(90%-100%)	2.14%	NA
5.	13q31.1	chr13	+	84869526	84877320	7795	86.84%	(80%-90%)	6.67%	NA
6.	4q33	chr4	-	170126345	170133883	7539	70.03%	(70%-80%)	0.79%	Insertic
7.	6p12.3	chr6	+	48873675	48879725	6051	64.84%	(60%-70%)	34.48%	Deletio
8.	Yp11.2	chrY	-	9273707	9279611	5905	59.83%	(50%-60%)	39.23%	Deletio
9.	8q24.3	chr8	+	145019974	145032719	12746	57.06%	(50%-60%)	0.79%	Insertic
10.	Yq11.223	chrY	+	21580120	21585551	5432	57.04%	(50%-60%)	38.30%	Deletio
11.	19q13.2	chr19	+	40954172	40959178	5007	56.66%	(50%-60%)	41.93%	Deletio
12.	Yp11.2	chrY	-	8121821	8126768	4948	54.57%	(50%-60%)	44.92%	Deletio
13.	Yp11.2	chrY	+	8996062	9000755	4694	50.80%	(50%-60%)	41.95%	Deletio
14.	Yq11.222	chrY	-	18622534	18626952	4419	47.33%	(40%-50%)	52.23%	Deletio
15.	Yq11.223	chrY	-	21845475	21850069	4595	43.18%	(40%-50%)	49.78%	Deletio /Inserti
16.	21q21.1	chr21	-	18563368	18566735	3368	34.26%	(30%-40%)	8.19%	NA
17.	5q33.3	chr5	-	156660448	156663815	3368	34.14%	(30%-40%)	8.50%	NA
18.	1q22	chr1	-	155629408	155632775	3368	33.75%	(30%-40%)	9.56%	NA
19.	7q36.1	chr7	-	150561277	150563994	2718	27.92%	(20%-30%)	10.27%	Deletio
20.	8q21.13	chr8	+	78652302	78654820	2519	26.60%	(20%-30%)	0.30%	NA
21.	10q24.2	chr10	-	99822511	99825532	3022	25.36%	(20%-30%)	24.65%	Deletio Insertic
22.	12q13.11	chr12	+	48509228	48511681	2454	18.44%	(10%-20%)	33.18%	Deletio Insertic
23.	Yq11.222	chrY	-	17669948	17671523	1576	17.11%	(10%-20%)	12.69%	Deletio

Next, the expected number of integration of HML-9 elements per chromosome was predicted and compared with the number of actually detected sites to assess whether HML-9 existed randomly in the human genome. The results showed that the number of HML-9 integration events observed is always inconsistent with the expected amounts (Fig. 1B-C). For the proviral elements, the number of HML-9 insertions in chromosomes 8, 13, 15, 16, 19, 21, and Y were higher than expected. In particular, the provirus elements in the Y chromosome were significantly higher than the predicted elements by the Chi-square test ($p < 0.05$). In chromosomes 1, 2, 4, 5, 6, 7, 10, and 12, the actual numbers identified are lower than expected (Fig. 1B). Especially, there is no provirus integration in chromosomes 3, 9, 11, 14, 17, 18, 20, 22, and X at all. With respect to the solo LTR elements, the number of HML-9 solo LTRs in chromosomes 2, 3, 14, 15, 18, 21, X, and Y were higher than expected. In chromosomes 1, 4, 5, 6, 7, 8, 10,11, 12, 13, 17, and 20 the actual numbers identified are lower than expected. Especially, there is no provirus integration in chromosomes 9, 16, 19, and 22 at all (Fig. 1C). Analysis revealed that HML-9 provirus and solo LTR integration displayed a non-randomly integration way among human chromosomes.

Further, all 23 identified provirus elements and 47 solo LTRs were analyzed to determine their locations in intergenic regions, intron, or exon (Table 1–2). The results show that 13 provirus elements are located in intergenic regions, accounting for 56.52%. 4 provirus elements are located in introns, accounting for 17.39%. 6 provirus elements are located in both introns and exons, accounting for 26.09% (Table 1). With respect to solo LTRs, 28 solo LTRs are located in intergenic regions, accounting for 59.57%. The remaining 19 solo LTRs are located in introns, accounting for 40.43% (Table 2). The results displayed an apparent insertion preference into intergenic regions and introns.

Table 2
HML-9 Solo LTR tracks distribution.

Number	Locus	Chromosome	Strand	Position start	Position end	Length (bp)	Percentage of LTR14C in length	Match + mismatch/full length	Range	Qgap(bp)/match + mismatch + Qgap(bp)
1	14q21.1	chr14	+	38011040	38012012	973	101.36%	6.91%	(0%-10%)	35.61%
2	Xq21.32	chrX	-	93273183	93274197	1015	100.85%	6.88%	(0%-10%)	2.47%
3	2q31.1	chr2	+	180236847	180237437	591	100.34%	6.84%	(0%-10%)	0.34%
4	18p11.31	chr18	-	4527618	4528209	592	100.17%	6.83%	(0%-10%)	0.00%
5	2q11.2	chr2	+	97964920	97965508	589	100.00%	6.82%	(0%-10%)	0.00%
6	15q14	chr15	+	39011033	39011621	589	100.00%	6.82%	(0%-10%)	0.00%
7	2p12	chr2	-	81304430	81305068	639	99.83%	6.81%	(0%-10%)	0.00%
8	2q32.3	chr2	-	194256159	194256746	588	99.83%	6.81%	(0%-10%)	0.00%
9	3q26.1	chr3	-	163283189	163283777	589	99.83%	6.81%	(0%-10%)	0.00%
10	4q26	chr4	+	116980222	116980809	588	99.83%	6.81%	(0%-10%)	0.34%
11	4p15.31	chr4	+	19556097	19556684	588	99.83%	6.81%	(0%-10%)	0.17%
12	7p21.2	chr7	+	14509240	14509827	588	99.83%	6.81%	(0%-10%)	0.00%
13	8q11.21	chr8	-	51178592	51179179	588	99.83%	6.81%	(0%-10%)	0.00%
14	11q12.3	chr11	-	62185237	62185824	588	99.83%	6.81%	(0%-10%)	0.00%
15	2q21.3	chr2	+	135521883	135522470	588	99.66%	6.80%	(0%-10%)	0.17%
16	3p12.2	chr3	-	81329902	81330488	587	99.66%	6.80%	(0%-10%)	0.17%
17	3p12.1	chr3	-	83618409	83618995	587	99.66%	6.80%	(0%-10%)	0.17%
18	5q21.3	chr5	-	105998962	105999549	588	99.66%	6.80%	(0%-10%)	0.34%
19	10p12.31	chr10	-	18856645	18857233	589	99.66%	6.80%	(0%-10%)	0.17%
20	Yq11.23	chrY	-	25974734	25975320	587	99.66%	6.80%	(0%-10%)	0.17%
21	6q14.1	chr6	-	82297755	82298498	744	99.49%	6.78%	(0%-10%)	0.34%
22	Xp22.2	chrX	+	11033746	11034330	585	99.32%	6.77%	(0%-10%)	0.51%
23	2p12	chr2	+	77807602	77808185	584	99.15%	6.76%	(0%-10%)	0.68%
24	1q23.3	chr1	+	162419359	162419942	584	98.98%	6.75%	(0%-10%)	0.17%
25	Xq27.2	chrX	-	142767872	142768454	583	98.98%	6.75%	(0%-10%)	0.00%
26	2q31.1	chr2	+	171365032	171365617	586	98.64%	6.73%	(0%-10%)	1.19%
27	5q13.3	chr5	+	75859521	75860102	582	98.64%	6.73%	(0%-10%)	0.17%
28	Xq27.3	chrX	-	144791258	144791846	589	98.47%	6.71%	(0%-10%)	1.37%
29	12q12	chr12	+	38144469	38145052	584	98.30%	6.70%	(0%-10%)	1.03%
30	15q21.3	chr15	-	54594796	54595373	578	98.13%	6.69%	(0%-10%)	2.04%
31	21q11.2	chr21	-	14080466	14081052	587	97.79%	6.67%	(0%-10%)	2.05%
32	4q28.2	chr4	+	129080872	129081454	583	97.61%	6.66%	(0%-10%)	2.22%
33	3q25.2	chr3	-	154944330	154944911	582	97.44%	6.64%	(0%-10%)	2.56%
34	11q24.2	chr11	+	124270705	124271275	571	96.93%	6.61%	(0%-10%)	0.00%
35	2q14.3	chr2	-	125024208	125024792	585	96.76%	6.60%	(0%-10%)	3.07%
36	6q27	chr6	+	169084226	169084808	583	96.76%	6.60%	(0%-10%)	3.07%
37	13q13.3	chr13	-	38319721	38320300	580	96.76%	6.60%	(0%-10%)	3.24%
38	7q35	chr7	+	143472173	143472744	572	96.08%	6.55%	(0%-10%)	3.75%
39	14q21.3	chr14	+	48011215	48011780	566	95.91%	6.54%	(0%-10%)	0.53%
40	3p21.31	chr3	+	44534488	44535059	572	95.74%	6.53%	(0%-10%)	3.77%

Number	Locus	Chromosome	Strand	Position start	Position end	Length (bp)	Percentage of LTR14C in length	Match + mismatch/full length	Range	Qgap(bp)/match + mismatch + Qgap(bp)
41	2q22.1	chr2	-	138860917	138861512	596	95.06%	6.48%	(0%-10%)	3.79%
42	12p13.32	chr12	-	4720007	4720593	587	94.89%	6.47%	(0%-10%)	4.95%
43	3p14.2	chr3	-	59469489	59470030	542	91.82%	6.26%	(0%-10%)	3.23%
44	1q24.2	chr1	+	168457190	168457732	543	90.80%	6.19%	(0%-10%)	0.37%
45	20p13	chr20	-	2809052	2809886	835	88.42%	6.03%	(0%-10%)	0.38%
46	18q21.33	chr18	+	63648105	63648555	451	76.49%	5.22%	(0%-10%)	0.22%
47	17q22	chr17	+	52961655	52962071	417	70.87%	4.83%	(0%-10%)	0.24%

2. Structural Characterization

To define the structural characteristics of HML-9 elements, the 23 proviruses were further analyzed by comparing them with the reference LTR14C-HERVK14C-LTR14C. According to the annotation information summarized in Dfam database (https://www.dfam.org/family/DF0000189/features), the complete HML-9 showed a typical proviral structure, containing 4 open reading frames (ORFs) and 2 flanking LTRs. In detail, the 5′ LTR located from nucleotide 1 to 587, the *gag* gene located from nucleotide 758 to 2548, the *pro* gene located from nucleotide 2548 to 3435, the *pol* gene located from nucleotides 3411 to 6060, the *env* gene located from nucleotides 5975 to 8020, and 3′ LTR located from nucleotide 8022 to 8608. We aligned the 23 HML-9 proviral sequences and annotated the position of the single retroviral component and deletions to describe the structure of each HML-9 provirus element (Fig. 2). Thereinto, the 16p12.3, 2p12, 15q21.1, 8p11.1, 13q31.1 and 4q33 are longer than 70% of complete reference in length. Further, the integrity of 6 separate regions (5′ LTR, *gag*, *pro*, *pol*, *env*, and 3′ LTR) was summarized in Table 3.

Table 3
The integrity of 6 separate regions relative to the corresponding sections of reference.

Number	Locus	Provirus Regions	5′LTR	gag	pro	pol	env	3′LTR
1	16p12.3	chr16 19393581 19402152	100.00%	99.83%	99.89%	99.39%	99.17%	99.66%
2	2p12	chr2 82022660 82031279	98.98%	99.72%	99.89%	99.43%	99.90%	99.15%
3	15q21.1	chr15 45234477 45243073	99.83%	99.27%	100.00%	99.66%	99.56%	99.83%
4	8p11.1	chr8 43694016 43702583	99.83%	99.44%	98.31%	99.39%	99.66%	99.83%
5	13q31.1	chr13 84869526 84877320	35.78%	99.55%	52.20%	99.77%	99.80%	99.32%
6	4q33	chr4 170126345 170133883	99.66%	99.89%	99.77%	99.70%	13.93%	0.00%
7	6p12.3	chr6 48873675 48879725	99.15%	92.79%	65.84%	6.13%	99.85%	99.66%
8	Yp11.2	chrY 9273707 9279611	88.42%	90.73%	63.81%	6.36%	98.83%	99.49%
9	8q24.3	chr8 145019974 145032719	0.00%	0.00%	0.79%	99.77%	99.90%	95.23%
10	Yq11.223	chrY 21580120 21585551	88.25%	91.74%	64.83%	6.25%	99.36%	15.84%
11	19q13.2	chr19 40954172 40959178	98.98%	98.94%	64.26%	6.17%	67.40%	77.51%
12	Yp11.2	chrY 8121821 8126768	99.66%	76.49%	14.09%	6.40%	99.17%	98.81%
13	Yp11.2	chrY 8996062 9000755	0.00%	80.23%	64.37%	6.47%	99.80%	95.55%
14	Yq11.222	chrY 18622534 18626952	96.76%	75.21%	0.00%	0.00%	84.75%	98.47%
15	Yq11.223	chrY 21845475 21850069	0.00%	70.85%	64.60%	6.40%	99.32%	99.15%
16	21q21.1	chr21 18563368 18566735	0.00%	0.00%	95.26%	96.12%	0.00%	0.00%
17	5q33.3	chr5 156660448 156663815	0.00%	0.00%	95.26%	96.12%	0.00%	0.00%
18	1q22	chr1 155629408 155632775	0.00%	0.00%	95.26%	96.12%	0.00%	0.00%
19	7q36.1	chr7 150561277 150563994	0.00%	0.00%	0.00%	16.91%	99.02%	54.51%
20	8q21.13	chr8 78652302 78654820	0.00%	0.00%	0.00%	0.00%	87.34%	100.00%
21	10q24.2	chr10 99822511 99825532	0.00%	0.00%	52.29%	95.43%	0.00%	0.00%
22	12q13.11	chr12 48509228 48511681	0.00%	0.00%	88.72%	63.52%	0.00%	0.00%
23	Yq11.222	chrY 17669948 17671523	52.81%	62.09%	0.00%	0.00%	0.00%	0.00%

It can be noted that many of the proviruses should be retrotransposed pseudogenes. Although these genes have been inactivated, they still can provide us with information about the evolutionary history of the gene family or the organism.

3. Phylogenetic analyses

To characterize the phylogenetic relationship of the HML-9 group, 5 proviral sequences longer than 80% of the HML-9 reference were screened to generate ML trees together with all representatives of Dfam HERV-K groups (HML-1 to 10) and exogenous Betaretroviruses. The results showed that the 5 identified proviruses all clustered with the Dfam HML-9 reference by a 100% bootstrap support (Fig. 3A). Subsequently, phylogenetic trees of a total of 44 identified solo LTRs longer than 90% of the LTR14C were constructed together with the LTR reference (Fig. 3B). Next, 4 ML trees of sub-regions whose lengths are longer than 90% of the corresponding section of the reference sequence were constructed, including 10 *gag* elements, 8 *pro* elements, 11 *pol* elements, and 13 *env* elements, respectively (Fig. 3C-E). These phylogenetic groups of different regions of HML-9 all clustered together and were clearly separated from the other HERV-K groups (HML1-8, 10). Thereinto, two distinct clusters in the *pro* and *pol* group were observed, respectively. They were statistically supported by 100% bootstrap values and were named type I and type II. The results showed that 21q21.1 (chr21:18563368–18566735), 1q22 (chr1:155629408–155632775), and 5q33.3 (chr5:156660448–156663815) were divided into type I; 15q21.1 (chr15:45234477–45243073), 16p12.3 (chr16:19393581–19402152), 4q33 (chr4: 170126345–170133883), 8p11.1 (chr8:43694016–43702583), and 2p12(chr2: 82022660–82031279) were divided into type II. Type II sequences included the Dfam HML-9 reference, whereas type I elements showed a more divergent relationship relative to the group references, which suggested these sequences were amplified at least 2 times after the initial genome integration.

4. Estimated time of integration

The HML-9 proviral age in the human genome was estimated based on the LTRs, *gag*, *pro*, *pol*, and *env* regions, respectively. Each region whose length exceeds 90% of the corresponding section of reference was selected to calculate the integration time. For *gag*, *pro*, *pol*, and *env* regions, the ancestral sequence of each region has been generated via Mega 7, following the ML method based on the multiple alignments of all elements. The T value (integration time) has been estimated by the relation T = D/0.2, where 0.2 represents the human genome neutral mutation rate expressed in substitutions/nucleotide/million years. To LTRs, as the 5' and 3' LTRs of the same provirus are identical at the time of integration and accumulate random substitutions in an independent manner [46], the T value was estimated by the relation T = D/0.2/2. For each region of a provirus, we provided details on the period of proviruses formation in Table 4. Overall, the time for evaluation based on the four regions (*gag*, *pro*, *pol*, and *env*) is larger than that for evaluation based on LTR. The majority of HML-9 elements (*gag*, *pro*, *pol*, and *env*) found in the human genome have been integrated between 37.5 and 151.5 million years ago (mya). The average time of integration was 76 mya. Interestingly, we found that the average integration time of type I was around 100 (99.92) mya, while the average integration time of type II is about 70 (68.85) mya. The integration time of type I was about 30 mya earlier than that of type II. However, the LTRs have been integrated between 17.5 and 48.5 million years ago (mya).

Table 4
Estimated time of HML-9 elements integration

locus	Provirus regions	Divergence from Consensus sequence				Mean Divergences	*T = D/0.2*	Age/ million years (gene vs consensus)	Divergence between 2 LTRs	T = D/0.2/2	Age/million years (LTR vs LTR)
		gag	pro	pol	env						
16p12.3	chr16 19393581 19402152	0.059	0.158	0.206	0.089	0.128	0.64	64.00	0.082	0.20500	20.50000
2p12	chr2 82022660 82031279	0.061	0.182	0.204	0.101	0.137	0.685	68.50	0.070	0.17500	17.50000
15q21.1	chr15 45234477 45243073	0.051	0.126	0.206	0.099	0.121	0.6025	60.25	0.080	0.20000	20.00000
8p11.1	chr8 43694016 43702583	0.091	0.177	0.231	0.121	0.155	0.775	77.50	0.107	0.26750	26.75000
13q31.1	chr13 84869526 84877320	0.054	NA	0.208	0.103	0.122	0.608333333	60.83	NA	NA	NA
4q33	chr4 170126345 170133883	0.058	0.172	0.214	NA	0.148	0.74	74.00	NA	NA	NA
6p12.3	chr6 48873675 48879725	0.063	NA	NA	0.106	0.085	0.4225	42.25	0.110	0.27500	27.50000
Yp11.2	chrY 9273707 9279611	0.114	NA	NA	0.139	0.127	0.6325	63.25	0.141	0.35250	35.25000
8q24.3	chr8 145019974 145032719	NA	NA	0.513	0.093	0.303	1.515	151.50	NA	NA	NA
Yq11.223	chrY 21580120 21585551	0.105	NA	NA	0.140	0.123	0.6125	61.25	NA	NA	NA
19q13.2	chr19 40954172 40959178	0.075	NA	NA		0.075	0.375	37.50	0.097	0.24250	24.25000
Yp11.2	chrY 8121821 8126768	NA	NA	NA	0.125	0.125	0.625	62.50	0.157	0.39250	39.25000
Yp11.2	chrY 8996062 9000755	NA	NA	NA	0.133	0.133	0.665	66.50	NA	NA	NA
Yq11.222	chrY:18622534–18626952	NA	NA	NA	NA	NA	NA	NA	0.194	0.48500	48.50000
Yq11.223	chrY 21845475 21850069	NA	NA	NA	0.143	0.143	0.715	71.50	NA	NA	NA
21q21.1	chr21 18563368 18566735	NA	0.266	0.190	NA	0.228	1.14	114.00	NA	NA	NA
5q33.3	chr5 156660448 156663815	NA	0.215	0.160	NA	0.188	0.9375	93.75	NA	NA	NA
1q22	chr1 155629408 155632775	NA	0.212	0.156	NA	0.184	0.92	92.00	NA	NA	NA
7q36.1	chr7 150561277 150563994	NA	NA	NA	0.197	0.197	0.985	98.50	NA	NA	NA
10q24.2	chr10 99822511 99825532	NA	NA	0.169	NA	0.169	0.845	84.50	NA	NA	NA

Additionally, the integration time of HML-9 elements in chimpanzees was also estimated. The results based on 4 internal regions (*gag*, *pro*, *pol*, and *env*) show they range from 22.5 mya to 117.5 mya (the average time of integration was 45.33 mya), which is later than that in humans (data not published).

5. Function prediction of cis-regulatory regions and enrichment analysis

The GREAT analysis can provide a prediction of potentially regulated genes based on spatial proximity. The results describing the associations between each solo LTR and its putatively regulated gene(s) were shown in Supplementary Table 1. There are a total of 69 genes that have been predicted. Among these, 5 solo LTRs are not associated with any genes, 15 solo LTRs are associated with 1 gene, and 27 solo LTRs are associated with 2 genes (Fig. 4A). The absolute distances to the transcription start site (TSS) of 3 genes are less than 5 kb, 13 genes are between 5 and 50 kb, 34 genes are between 50 and 500 kb, and 19 genes are more than 500 kb (Fig. 4B-C).

To analyze the biological classification of key genes related to solo LTRs, GO Slim summaries were performed. The summary of biological processes (BP) revealed that these genes were mainly enriched in biological regulation, metabolic process, multicellular organismal process, response to stimulus, cell communication, developmental process, localization, and cellular component organization (Fig. 4D). The GO Slim summary of cellular component (CC) showed that these genes significantly enriched in membrane and nucleus, and GO Slim summary of the molecular function (MF) showed that these genes significantly enriched in protein binding and ion binding (Fig. 4E-F).

Further, these potential regulatory genes are all annotated to the selected functional categories and used for enrichment analysis. According to the false discovery rate (FDR) value, the TOP 10 significant GO terms for biological process included the regulation of endothelial cell chemotaxis, regulation of natural killer cell-mediated immunity, positive regulation of synapse assembly, natural killer cell-mediated immunity, regulation of synapse assembly, positive regulation of chemotaxis, synapse assembly, regulation of synapse organization, regulation of synapse structure or activity, and synapse organization (Fig. 5A). The bar chart shows the enrichment ratio of the results. When top results are chosen to be returned and the FDR for the categories is ≤ 0.05, the colors of the bars are in a darker shade than when the FDR exceeds 0.05 (Fig. 5A). The volcano plot in Fig. 5B shows the log2 of the FDR versus the enrichment ratio for all the functional categories in the database, highlighting the degree by which the significant categories stand out from the background. The size and color of the dot are proportional to the number of overlapping (for ORA). The significantly enriched categories are labeled, and the labels are positioned automatically by a force field-based algorithm at startup. The enrichment results for cellular component and molecular function are illustrated in Fig. 5C-F. It must be noted that these results are entirely predictive and that future research is required to confirm any of the implied associations between the solo LTRs and the nearby genes.

In addition to solo LTRs, the prediction of putatively regulated genes by proviral LTRs was also performed by GREAT. Enrichment analysis was carried out as described for solo LTR. The results describing the associations between each proviral LTR and its putatively regulated gene(s) were shown in Supplementary Table 2. There are a total of 36 genes that have been predicted. Among these, 4 proviral LTRs are not associated with any genes, 6 proviral LTRs are associated with 1 gene, and 15 proviral LTRs are associated with 2 genes (Supplementary Fig. 1A). Unexpectedly, the 4 proviral LTRs associated with none genes belong to two pairs of 5’ and 3’ LTRs of the same provirus, 2p12 (chr2 82022660 82031279) and Yp11.2 (chrY 8121821 8126768), respectively. Especially, 2p12 (chr2 82022660 82031279) is a rather complete provirus. The absolute distances to the TSS of 0 genes are less than 5 kb, 13 genes are between 5 and 50 kb, 15 genes are between 50 and 500 kb, and 8 genes are more than 500 kb (Supplementary Fig. 1B-C). The biological classification of key genes related to provirus LTRs by GO Slim summaries was shown in Supplementary Fig. 1D-F). The enrichment results for biological process, cellular component, and molecular function were illustrated in Supplementary Fig. 2–4.

6. In silico examination of the conserved transcription factor binding sites

HML-9 exhibiting specific base insertions may influence the complexity of LTR transcriptional regulation [51]. A complete view of the putative transcription factor binding sites within the HML-9 LTR were shown in Fig. 6A. A total of 22 human transcription binding sites were predicted, including 19 transcription factors: EHF, SOX10, FOS, FOSL1, FOSL2, JUNB, JUND, ETV4, KLF1, KLF5, KLF7, ZNF263, THAP1, SP4, RBPJ, HAND2, MAZ, NEUROG2, and NEUROD1. The motifs are marked on the sense strand and antisense strand of the consensus sequence.

7. PBS type of HML-9 sequences

Traditionally HERVs have been named per the tRNA that binds their RT enzyme and PBS [52]. Thus, HERV-K is named after lysine-tRNA. In the analyzed 5 proviral and consensus sequences of HML-9 elements, the PBS was located approximately from nucleotide 3 to 20 in nucleotides downstream the 5’LTR. To summarize the general variation of the PBS sequence within the HML-9 group we generated a logo in which the letter height is proportional to the nucleotides/amino acid conservation at each position (Fig. 6B-C). The result showed the TGG starting nucleotides were the most conserved among the 18 bases analyzed which had a tryptophan (W) PBS type, followed by arginine (R), confirming the relatively low taxonomic value of this feature.

Discussion

Until now, the characterization and identification of HML1-8, and HML10 groups have been carried out [9, 42, 51, 53–58]. However, there are only very few incomplete characterizations and research for HML-9 at present. Previous research revealed that HERVK14C (HML-9) could be regulated by KAP1 in human cells and contribute to innate immune. KAP1 represses HERVK14C and ZNF genes, both of which overlap with KAP1 binding sites and H3K9me3 in multiple cell types [39]. The study on HML-9 expression is not enough to reveal the true role of these elements that play in the genome. Therefore, it is necessary to carry out a systematic and comprehensive characterization of the HML-9.

Our research followed the approach carried out in previously published studies [9, 57], and mapped out the HML-9 proviruses and solo LTRs information completely in the human genome, thus providing an exhaustive characterization of this group including genomic distribution, structural characterization, phylogenesis, integration time and regulatory function prediction. A total of 23 HERV-K HML-9 provirus and 47 solo LTR elements have been characterized. The chromosomal distribution of these proviruses and the solo LTRs revealed a non-random integration pattern. The results display an apparent insertion preference into intergenic regions and introns. Especially, the provirus in the Y chromosome was significantly different from the predicted sequences by the Chi-square test ($p = 0.01$), which indicated the MSY accumulate higher densities of HERVs and the associated sequences [24].

Phylogenetic analyses showed that 5 sequences of HML-9 near-full-length proviruses as well as 10 elements of *gag*, 8 elements of *pro*, 11 elements of *pol*, and 13 sequences of *env* formed a unique monophyletic cluster and were supported by the maximum bootstrap value, clearly divided from other HML groups. The phylogenetic trees of *pro* and *pol* regions both revealed the presence of two well-supported clusters, identified here as type I and II, which were statistically

supported by 100 bootstrap values. Type II cluster included the Dfam HML-9 reference, whereas type I cluster showed a more divergent structure relative to the group references, suggesting these sequences might experience 2 genome integrations. In addition, the integration time of HML-8 proviruses was calculated using the LTR, *gag*, *pro*, *pol*, and *env* regions. The results indicated that the main period of HML-9 integration based on 4 internal regions is between 37.5 and 151.5 mya. The average time of integration was 76 mya. The integration time of type I was around 100 mya, while the integration time of type II is about 70 mya. The integration time of type I was about 30 mya earlier than that of type II.

Further, we performed prediction and cluster analysis of potential regulatory genes for both HML-9 *pro*virus and solo LTRs. There are a total of 69 genes have been predicted. The biological process and molecular function analysis showed these genes associated with synapses. Previous studies have shown that the HERV-W can interfere with neuronal protrusions, alter the N-methyl-d-aspartate receptor (NMDAR)-mediated synaptic organization and plasticity through glia- and cytokine-dependent changes [59]. Here our work suggested that the HML-9 LTRs regulated genes may also be widely involved in the function of synapses. Furthermore, the prediction of transcription factors on HML-9 elements by JASPAR also indicated that HML-9 is likely to play a regulatory role in downstream genes. In addition, for the analysis of PBS of HML-9, we confirmed the conclusion that this nomenclature is imprecise because HML-9 belongs to the HERV-K subgroup but its PBS analysis results were tryptophan nevertheless. It should be noted that these results are entirely speculative. And experimental validation studies are required to confirm the associations between the solo LTRs and these genes.

Conclusion

There is a previous study of HML-9 (HERVK14C) indicating that HML-9 could exert in different tissues both in physiological conditions as well as involvement in human disease development, possibly contributing to immune regulation and antiviral defense. In order to systematically and clearly study the important role of HML-9 in pathological and physiological processes, the current work detailed a clear description of all HML-9 elements integrated into the human genome which could contribute to better defining their real impact and contribution to our genome.

Abbreviations

HERVs, Human endogenous retroviruses

TEs, transposable elements

REs, transcription factors

LTRs, long terminal repeats

PBS, primer binding site

HML, human MMTV-like

MSY, male-specific regions of the Y chromosome

RT, reverse transcriptase

ML, maximum likelihood

NNI, nearest neighbor interchange

GREAT, Genomic Regions Enrichment of Annotations Tool

WebGestalt, WEB-based Gene SeT AnaLysis Toolkit

ORA, Over-Representation Analysis

GSEA, Gene Set Enrichment Analysis

ORFs, open reading frames

Mya, million years ago

TSS, transcription start site

BP, biological processes

CC, cellular component

MF, molecular function

FDR, false discovery rate

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analyzed during this study are included in this published article.

Competing interests

The authors declare that they have no competing interests.

Funding information

This study was supported by the NSFC (31900157).

Author Contributions

Research design: L.L and Y.C. Performed the analysis: L.J, J.H, H.L, X.Z, X.W, Y.L, T.L, Z.B, and J.L. Contributed to the composition of the manuscript: L.J, M.L, and L.L.

Acknowledgements

This study was supported by the NSFC (31900157).

References

1. Griffiths DJ: **Endogenous retroviruses in the human genome sequence.** *Genome Biol*2001, **2:**REVIEWS1017.
2. Kazazian HH, Jr.: **Mobile elements: drivers of genome evolution.** *Science*2004, **303:**1626-1632.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: **Initial sequencing and analysis of the human genome.** *Nature*2001, **409:**860-921.
4. Gogvadze E, Buzdin A: **Retroelements and their impact on genome evolution and functioning.** *Cellular and molecular life sciences : CMLS*2009, **66:**3727-3742.
5. Bannert N, Kurth R: **The evolutionary dynamics of human endogenous retroviral families.** *Annu Rev Genomics Hum Genet*2006, **7:**149-173.
6. Bannert N, Kurth R: **Retroelements and the human genome: new perspectives on an old relation.** *Proc Natl Acad Sci U S A*2004, **101 Suppl 2:**14572-14579.
7. Boller K, Schönfeld K, Lischer S, Fischer N, Hoffmann A, Kurth R, Tönjes RR: **Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles.** *The Journal of general virology*2008, **89:**567-572.
8. Schiavetti F, Thonnard J, Colau D, Boon T, Coulie PG: **A human endogenous retroviral sequence encoding an antigen recognized on melanoma by cytolytic T lymphocytes.** *Cancer research*2002, **62:**5510-5516.
9. Grandi N, Cadeddu M, Pisano MP, 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.** *Mobile DNA*2017, **8:**15.
10. Medstrand P, van de Lagemaat LN, Mager DL: **Retroelement distributions in the human genome: variations associated with age and proximity to genes.** *Genome research*2002, **12:**1483-1495.
11. Andersson ML, Lindeskog M, Medstrand P, Westley B, May F, Blomberg J: **Diversity of human endogenous retrovirus class II-like sequences.** *J Gen Virol*1999, **80 (Pt 1):**255-260.
12. Sverdlov ED: **Retroviruses and primate evolution.** *BioEssays : news and reviews in molecular, cellular and developmental biology*2000, **22:**161-171.
13. Belshaw R, Dawson ALA, Woolven-Allen J, Redding J, Burt A, Tristem M: **Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity.** *Journal of virology*2005, **79:**12507-12514.
14. Mayer J, Sauter M, Rácz A, Scherer D, Mueller-Lantzsch N, Meese E: **An almost-intact human endogenous retrovirus K on human chromosome 7.** *Nature genetics*1999, **21:**257-258.
15. Hughes JF, Coffin JM: **Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome.** *Genetics*2005, **171:**1183-1194.

16. Medstrand P, Mager DL: **Human-specific integrations of the HERV-K endogenous retrovirus family.** *Journal of virology*1998, **72:**9782-9787.
17. Costas J: **Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length proviral genomes.** *Journal of molecular evolution*2001, **53:**237-243.
18. Smit AF: **Interspersed repeats and other mementos of transposable elements in mammalian genomes.** *Current opinion in genetics & development*1999, **9:**657-663.
19. van de Lagemaat LN, Medstrand P, Mager DL: **Multiple effects govern endogenous retrovirus survival patterns in human gene introns.** *Genome biology*2006, **7:**R86.
20. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL: **Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line.** *PLoS genetics*2006, **2:**e2.
21. Cutter AD, Good JM, Pappas CT, Saunders MA, Starrett DM, Wheeler TJ: **Transposable element orientation bias in the Drosophila melanogaster genome.** *J Mol Evol*2005, **61:**733-741.
22. Medstrand P, van de Lagemaat LN, Mager DL: **Retroelement distributions in the human genome: variations associated with age and proximity to genes.** *Genome Res*2002, **12:**1483-1495.
23. Smit AF: **Interspersed repeats and other mementos of transposable elements in mammalian genomes.** *Curr Opin Genet Dev*1999, **9:**657-663.
24. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, et al: **The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.** *Nature*2003, **423:**825-837.
25. Zhdanov VM, Soloviev VD, Bektemirov TA, Ilyin KV, Bykovsky AF, Mazurenko NP, Irlin IS, Yershov FI: **Isolation of oncornaviruses from continuous human cell cultures.** *Intervirology*1973, **1:**19-26.
26. Sarngadharan MG, Sarin PS, Reitz MS, Gallo RC: **Reverse transcriptase activity of human acute leukaemic cells: purification of the enzyme, response to AMV 70S RNA, and characterization of the DNA product.** *Nat New Biol*1972, **240:**67-72.
27. Matteucci C, Balestrieri E, Argaw-Denboba A, Sinibaldi-Vallebona P: **Human endogenous retroviruses role in cancer cell stemness.** *Semin Cancer Biol*2018, **53:**17-30.
28. Barth M, Groger V, Cynis H, Staege MS: **Identification of human endogenous retrovirus transcripts in Hodgkin Lymphoma cells.** *Mol Biol Rep*2019, **46:**1885-1893.
29. Grabski DF, Hu Y, Sharma M, Rasmussen SK: **Close to the Bedside: A Systematic Review of Endogenous Retroviruses and Their Impact in Oncology.** *J Surg Res*2019, **240:**145-155.
30. Cañadas I, Thummalapalli R, Kim JW, Kitajima S, Jenkins RW, Christensen CL, Campisi M, Kuang Y, Zhang Y, Gjini E, et al: **Tumor innate immunity primed by specific interferon-stimulated endogenous retroviruses.** *Nature medicine*2018, **24:**1143-1150.
31. Arru G, Mameli G, Deiana GA, Rassu AL, Piredda R, Sechi E, Caggiu E, Bo M, Nako E, Urso D, et al: **Humoral immunity response to human endogenous retroviruses K/W differentiates between amyotrophic lateral sclerosis and other neurological diseases.** *European journal of neurology*2018, **25:**1076-1071e1084.
32. Mameli G, Erre GL, Caggiu E, Mura S, Cossu D, Bo M, Cadoni ML, Piras A, Mundula N, Colombo E, et al: **Identification of a HERV-K env surface peptide highly recognized in Rheumatoid Arthritis (RA) patients: a cross-sectional case-control study.** *Clinical and experimental immunology*2017, **189:**127-131.
33. Arru G, Galleri G, Deiana GA, Zarbo IR, Sechi E, Bo M, Cadoni MPL, Corda DG, Frau C, Simula ER, et al: **HERV-K Modulates the Immune Response in ALS Patients.** *Microorganisms*2021, **9**.
34. Xue B, Zeng T, Jia L, Yang D, Lin SL, Sechi LA, Kelvin DJ: **Identification of the distribution of human endogenous retroviruses K (HML-2) by PCR-based target enrichment sequencing.** *Retrovirology*2020, **17:**10.
35. Denne M, Sauter M, Armbruester V, Licht JD, Roemer K, Mueller-Lantzsch N: **Physical and functional interactions of human endogenous retrovirus proteins Np9 and rec with the promyelocytic leukemia zinc finger protein.** *Journal of virology*2007, **81:**5607-5616.
36. Buscher K, Hahn S, Hofmann M, Trefzer U, Ozel M, Sterry W, Lower J, Lower R, Kurth R, Denner J: **Expression of the human endogenous retrovirus-K transmembrane envelope, Rec and Np9 proteins in melanomas and melanoma cell lines.** *Melanoma Res*2006, **16:**223-234.
37. Kraus B, Fischer K, Büchner SM, Wels WS, Löwer R, Sliva K, Schnierle BS: **Vaccination directed against the human endogenous retrovirus-K envelope protein inhibits tumor growth in a murine model system.** *PloS one*2013, **8:**e72756.
38. Antony JM, van Marle G, Opii W, Butterfield DA, Mallet F, Yong VW, Wallace JL, Deacon RM, Warren K, Power C: **Human endogenous retrovirus glycoprotein-mediated induction of redox reactants causes oligodendrocyte death and demyelination.** *Nat Neurosci*2004, **7:**1088-1095.
39. Tie CH, Fernandes L, Conde L, Robbez-Masson L, Sumner RP, Peacock T, Rodriguez-Plata MT, Mickute G, Gifford R, Towers GJ, et al: **KAP1 regulates endogenous retroviruses in adult human cells and contributes to innate immune control.** *EMBO Rep*2018, **19**.
40. Kent WJ: **BLAT–the BLAST-like alignment tool.** *Genome research*2002, **12:**656-664.
41. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome research*2002, **12**.
42. Grandi N, Pisano MP, Pessiu E, Scognamiglio S, Tramontano E: **HERV-K(HML7) Integrations in the Human Genome: Comprehensive Characterization and Comparative Analysis in Non-Human Primates.** *Biology (Basel)*2021, **10**.
43. Hall TA: **BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. .** *Nuclei Acids***41:95-98.** .
44. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.** *Mol Biol Evol*2016, **33:**1870-1874.
45. Letunic I, Bork P: **Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation.** *Nucleic Acids Res*2021, **49:**W293-W296.

46. Lebedev YB, Belonovitch OS, Zybrova NV, Khil PP, Kurdyukov SG, Vinogradova TV, Hunsmann G, Sverdlov ED: **Differences in HERV-K LTR insertions in orthologous loci of humans and great apes.** *Gene*2000, **247:**265-277.
47. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G: **GREAT improves functional interpretation of cis-regulatory regions.** *Nature biotechnology*2010, **28:**495-501.
48. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B: **WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs.** *Nucleic Acids Res*2019, **47:**W199-W205.
49. 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.
50. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome Res*2004, **14:**1188-1190.
51. Subramanian RP, Wildschutte JH, Russo C, Coffin JM: **Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses.** *Retrovirology*2011, **8:**90.
52. Cohen M, Larsson E: **Human endogenous retroviruses.** *Bioessays*1988, **9:**191-196.
53. 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.
54. Mayer J, Meese EU: **The human endogenous retrovirus family HERV-K(HML-3).** *Genomics*2002, **80:**331-343.
55. Seifarth W, Baust C, Murr A, Skladny H, Krieg-Schneider F, Blusch J, Werner T, Hehlmann R, Leib-Mosch C: **Proviral structure, chromosomal location, and expression of HERV-K-T47D, a novel human endogenous retrovirus derived from T47D particles.** *J Virol*1998, **72:**8384-8391.
56. 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.
57. Pisano MP, 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**.
58. Liu M, Jia L, Li H, Liu Y, Han J, Zhai X, Wang X, Li T, Li J, Zhang B, et al: **Identification and characterization of the HERV-K (HML-8) group of human endogenous retroviruses in the genome.** *bioRxiv*2022**:**2022.2002.2010.479833.
59. Johansson EM, Bouchet D, Tamouza R, Ellul P, Morr AS, Avignone E, Germi R, Leboyer M, Perron H, Groc L: **Human endogenous retroviral protein triggers deficit in glutamate synapse maturation and behaviors associated with psychosis.** *Science advances*2020, **6:**eabc0708.

Figures

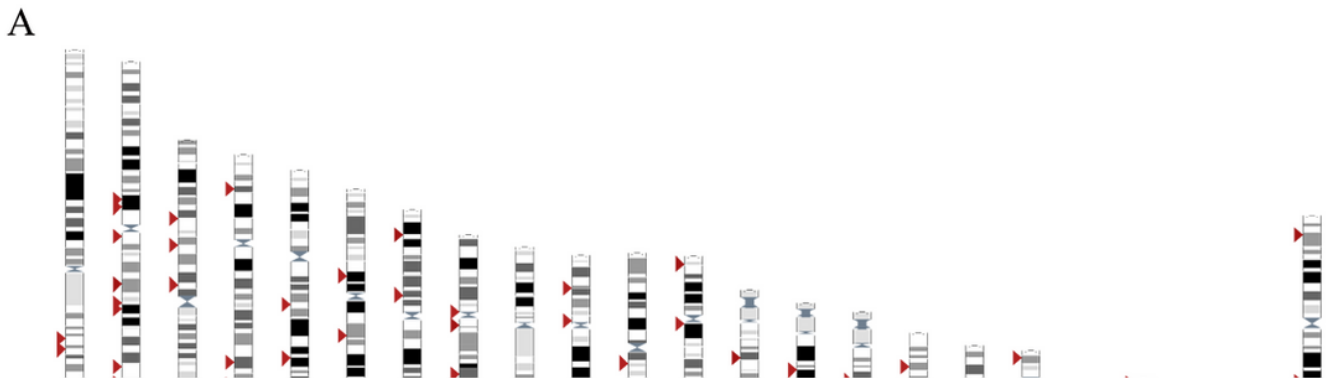

Figure 1

Chromosomal distribution of HML-9 loci. (A) All HML-9 elements (red arrows) have been visualized on the human karyotype (www.ensembl.org). The number of HML-9 proviral elements **(B)** and solo LTRs **(C)** integrated into each human chromosome was described and compared to the expected number of insertion events. The expected number of sequences in each chromosome is marked in blue and the actual number of sequences detected is marked in orange.

Figure 2

HML-9 proviruses structural characterization. Each HML-9 provirus element has been detailed and compared to the Dfam reference sequence. LTRs, *gag*, *pro*, *pol*, and *env* genes were annotated. Black lines represented deleted regions.

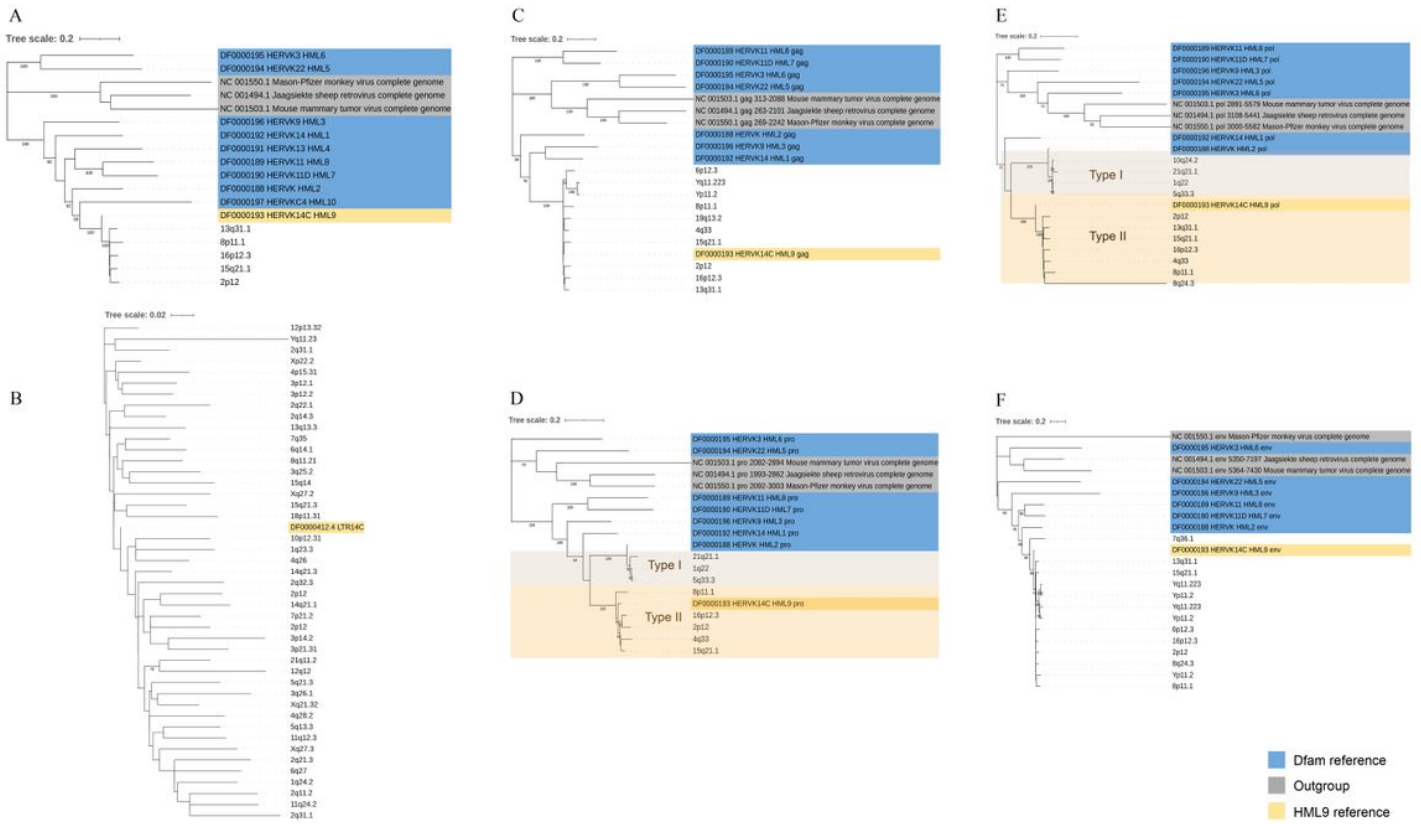

Figure 3

Phylogenetic analysis of the HML-9 near-full-length proviruses, solo LTRs, and 4 subregions by Maximum Likelihood method. Phylogenetic analyses of 5 HML-9 proviruses elements **(A)**, 44 solo LTRs **(B)**, 10 *gag* elements **(C)**, 8 *pro* elements **(D)**, 11 *pol* elements **(E)**, and 13 *env* elements **(F)** together with references. The two intragroup clusters of the *pro* and *pol* genes (type I and II) were annotated and depicted with brown and orange background colors, respectively. The resulting phylogeny was tested by the bootstrap method with 500 replicates. Length of branches indicates the number of substitutions per site.

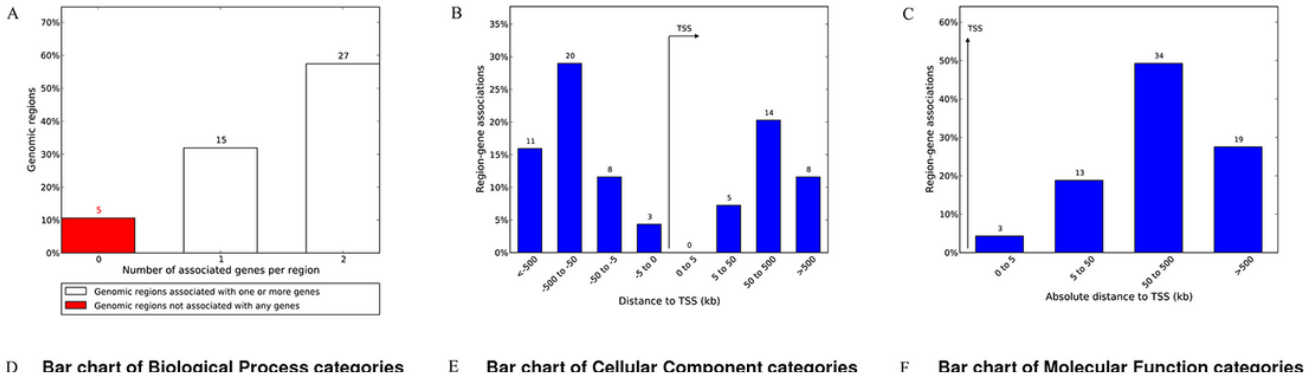

Figure 4

The genes associated with solo LTRs and GO Slim summary.

(A) The number of associated genes per solo LTR. **(B)** Binned by orientation and distance to TSS. **(C)** Binned by absolute distance to TSS. Each biological process summary **(D)**, cellular component summary **(E),** and molecular function summary **(F)** are represented by a red, blue, and green bar, respectively. The height of the bar represents the number of IDs in the gene list and also in the category.

Figure 5

Enrichment result categories binned by biological process, cellular component, and molecular function. (A) and **(B),** The bar chart plots the enrichment results and customizable volcano plot of the biological process. **(C)** and **(D),** The bar chart plots the enrichment results and customizable volcano plot of the cellular component. **(E)** and **(F),** The bar chart plots the enrichment results and customizable volcano plot of molecular function.

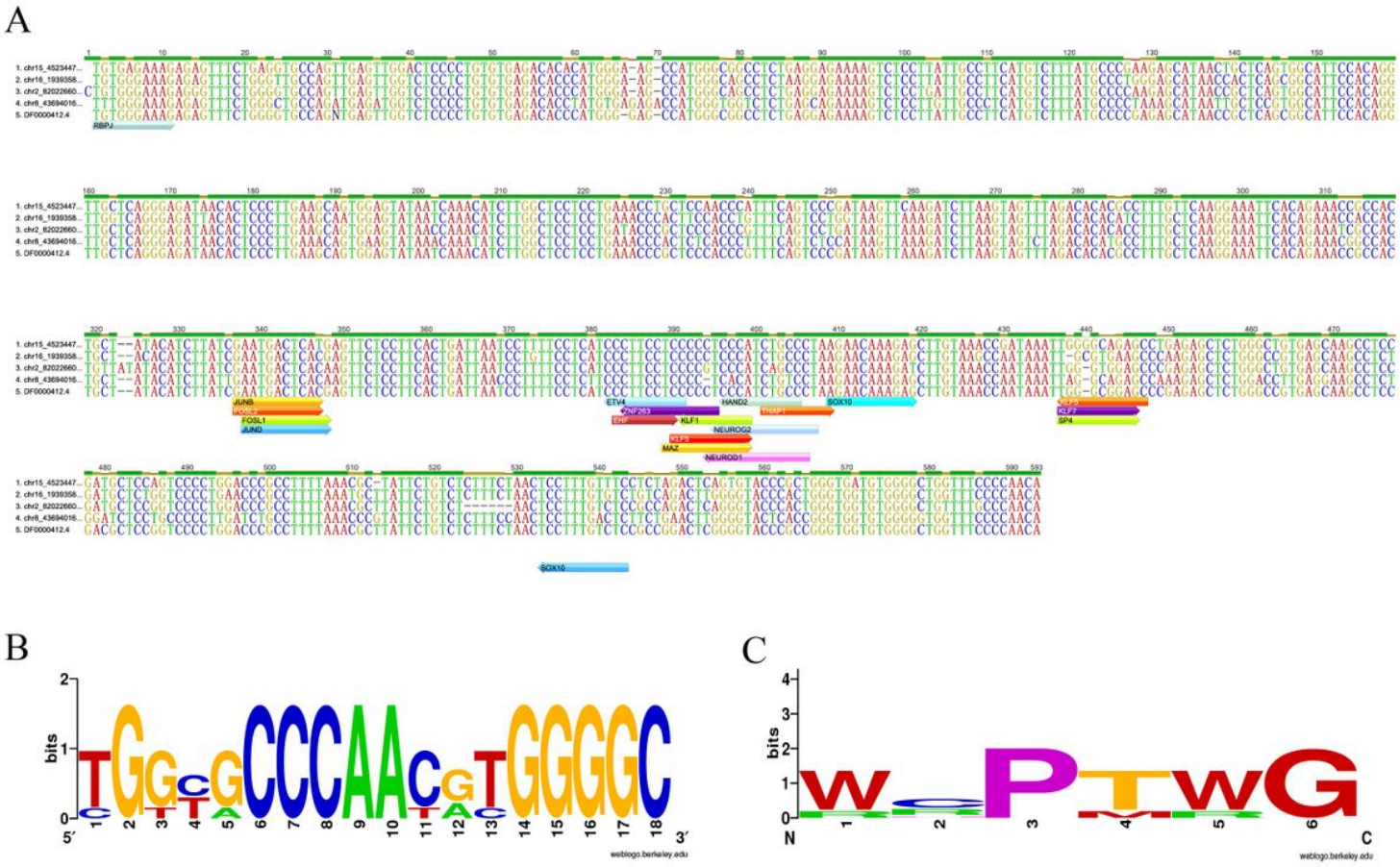

Figure 6

In silico examination of the conserved transcription factor binding sites and logos representing the PBS of HML-9. (A) The forward arrows indicate sense strand, the reverse arrow indicates antisense strand. Different transcription factors are marked with different colors. **(B)** PBS nucleotide sequence; **(C)** PBS amino acid sequence. Created at http://weblogo.berkeley.edu/logo.cgi

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigure1.tif
- SupplementaryFigure2.tif
- SupplementaryFigure3.tif
- SupplementaryFigure4.tif
- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- Supplementarydataset1HML9proviruselements.fas
- Supplementarydataset2HML9soloLTRelements.fas