J Mol Evol (1999) 48:102–111

JOURNAL OF MOLECULAR EVOLUTION

Biased Nucleotide Composition of the Genome of HERV-K Related Endogenous Retroviruses and Its Evolutionary Implications

J. Zsíros,[1,2] M.F. Jebbink,[1] V.V. Lukashov,[1] P.A. Voûte,[2] B. Berkhout[1]

[1] Department of Human Retrovirology, Academic Medical Center, University of Amsterdam, Meibergdreef 15, 1105 AZ Amsterdam, The Netherlands
[2] Department of Pediatric Oncology, Academic Medical Center, University of Amsterdam, Meibergdreef 15, 1105 AZ Amsterdam, The Netherlands

Received: 9 February 1998 / Accepted: 8 June 1998

Abstract. The human genome contains a large number of sequences that belong to the HERV-K family of human endogenous retroviruses. Most of these elements are likely remnants of ancient infections by ancestral exogenous retroviruses. To obtain further insight into the evolutionary history and molecular mechanisms responsible for the diversity of the human HERV-K elements, we analyzed several aspects of their genome structure. The nucleotide composition of the HERV-K genome was found to be highly biased and asymmetric, with an abundance of the A nucleotide in the viral (+) strand. A similar trend has been reported for the genomes of several exogenous retroviruses, with different nucleotides as the preferred building block. Other genome characteristics that were reported previously for actively replicating retroviruses are also apparent for the endogenous HERV-K virus. In particular, we observed suppression of the dinucleotide CpG, which represents potential methylation sites, and a strong preference for synonymous substitutions within the open reading frame of the reverse transcriptase (RT) enzyme. Furthermore, the mutational spectrum of the HERV-K RT enzyme was evaluated by nucleotide sequence comparison of 34 available elements. Interestingly, this analysis revealed a striking similarity with the mutational pattern of the HIV-1 RT enzyme, with a preference for G-to-A and C-to-T transitions. It is proposed that the mutational bias of the HERV-K RT enzyme played a role in the shaping of this retroviral genome, which was actively replicating more than 30 million years ago. This effect can still be observed in the contemporary endogenous HERV-K elements.

Key words: Human endogenous retroviruses — HERV-K10 family — Retrovirus evolution — Nucleotide composition — Mutational bias

Introduction

The genomes of humans and other primates contain several families of related sequences, termed human endogenous retroviruses (HERVs), that resemble infectious retroviruses (reviewed by Wilkinson et al. 1994; Lower et al. 1996). These elements are remnants of germline infections of the human ancestor by ancient retroviruses. Some members of the current phylogeny may have arisen subsequently by intracellular retrotransposition. The HERV-K family of endogenous elements inserted into the primate genome after the divergence of the New World monkeys, which occurred about 30 million years ago (Steinhuber et al. 1995). This family has attracted much attention for several reasons. First, the HERV-K family is relatively extended, containing at least 55 members per haploid human genome (Zsiros et al. 1998b) that can be grouped in six subfamilies with a nucleotide sequence dissimilarity of about 25% between the groups (Medstrand and Blomberg 1993). This current HERV-K

Correspondence to: B. Berkhout; *e-mail:* B.Berkhout@AMC.UVA.NL

phylogeny likely represents independent introductions of related exogenous retroviruses into the human genome. Second, several HERV-K elements are transcriptionally active (Medstrand et al. 1992; Andersson et al. 1996), and differential expression of HERV-K genes has been reported in different tumor samples (Li et al. 1995). Third, reading frames for the major gene products Gag, Pol, and Env appear to be well preserved, at least in some members of the HERV-K family (Ono et al. 1986; Mueller-Lantzsch et al. 1993; Lower et al. 1993, 1996). This is in striking contrast with many other endogenous retroviral families, in which the most viral genes are severely disrupted by numerous mutations. Fourth, HERV-K elements were recently found to be associated with the virion particles which were detected in placenta material and which contain an enzymatically active reverse transcriptase (RT) protein (Simpson et al. 1996; Patience et al. 1996). Despite these experimental findings, it is not known whether HERV-K elements are currently still active either in intracellular retrotransposition or infection.

HERV-K10 is the prototype of the HERV-K family of human endogenous retroviral elements and was detected and fully sequenced in 1986 (Ono et al. 1986). This family shows approximately 60% nucleotide (nt) sequence similarity with the mouse mammary tumor virus (MMTV). Additional members of the family were identified more recently and clustered into six groups, termed HML-1 to HML-6 (HML = human endogenous MMTV-like sequences) (Li et al. 1996; Medstrand and Blomberg 1993; Medstrand et al. 1997; Zsíros et al. 1998a,b). So far, an approximately 600-nt fragment of the RT gene has been sequenced for 34 HERV-K members, allowing a detailed phylogenetic analysis of this retrovirus family (Zsíros et al. 1998a,b).

In order to gain further insight into the evolutionary history of these endogenous elements, we performed an analysis of the nucleotide composition of their genomes. Here we report that the nucleotide composition of these elements is strikingly nonrandom. In particular, all HERV-K genomes contain an abundance of the A nucleotide (27.9–37.5%) in the coding (+) strand, which appears to be counteracted by suppression of the C nucleotide. We also observed a bias against CpG methylation sites and a strong preference for silent codon changes within the open reading frames (orf). Furthermore, comparison of the sequences of different HERV-K members allowed us to estimate the mutational spectrum of the HERV-K RT enzyme, which is remarkably similar to that of the HIV-1 RT enzyme. To our knowledge, this is the first report of a biased nucleotide content of endogenous retroviral genomes. We discuss the evolutionary constraints that may have been acting on the precursor exogenous retrovirus that existed some 30 million years ago.

Materials and Methods

Sequence Analysis. For sequence analysis we used an approximately 600-nt fragment of the RT gene of the HERV-K-related genomes. This fragment corresponds to the region between position 4099 and position 4693 in the nucleotide sequence of the prototype HERV-K10 element. Sequences of 34 HERV-K-related RT segments were available. Most of them were obtained previously by (RT-)PCR amplification of cellular RNA by our group or by others. For the exogenous retroviruses, the corresponding region was identified by nucleotide and amino acid sequence alignment.

Nucleotide sequence alignments and analyses were performed with PC/GENE software (IntelliGenetics) and appropriate programs of the GCG Sequence Analysis Software Package (Genetics Computer Group). Sequence alignments were optimized manually. Positions with an alignment gap were excluded from the analysis. Synonymous and nonsynonymous nucleotide p-distances (d_s and d_a, respectively) between two sequences were calculated with the MEGA program using the Nei–Gojobori method. The d_s/d_a ratio for a group of sequences was calculated as described previously (Lukashov et al. 1995) according to the formula $d_s/d_a = (\sum M_{si}/\sum S_{si})/(\sum M_{ai}/\sum S_{ai})$, where M_{si} and M_{ai} represent the numbers of mutation events at coding synonymous and nonsynonymous sites, respectively, and S_{si} and S_{ai} the numbers of coding synonymous and nonsynonymous sites, respectively.

The mutational spectrum of the HERV-K elements was evaluated by analysis of the interspecies nucleotide substitutions within the HML subgroups. We used the subgroup consensus sequence as prototype sequence and scored the sequence changes in individual members. The analysis was restricted to subgroups with more than one member. When identical nucleotide changes were observed in multiple members, they were counted only once. No correction was made for the nucleotide composition of the RT gene fragment.

Sources and GenBank Accession Numbers. The nucleotide sequences of human exogenous and endogenous retroviruses analyzed in this study are as as follows: clones M3.8 (accession number: U87587), HP.1 (U87588), K1.1 (U87589), N8.4 (U87590), P1.1 (U87591), M3.5 (U87592), P1.8 (U87593), P1.10 (U87594), D1.2 (U87595), D1.3 (U87596), L4.4, H3.1, St.2, and P1.4 from Zsíros et al. (1988a); clones St.4 (AF030038), P1.6 (AF030039), P1.7 (AF030040), N5.1 (AF030041), M3.1 (AF030042), P1.3 (AF030043), HP.2 (AF030044), M3.9 (AF030045), M3.10 (AF030046), and St.1 (AF030047) from Zsíros et al. (1998b); HERV-K10 (M14123) from Ono et al. (1986); T47D (U47118) from Patience et al. (1996); HERV-(k)27, HERV-(k)67, and HERV-(K)73 from Li et al. (1996); HERV-(K)55 (U39936) from Li et al. (1995); HM16 (M30520) from Deen and Sweet (1986); ERV MLN (U27242) from Seifarth et al. (1995); HERV-76 and HERV-50 from Li et al. (1996); human immunodeficiency virus type 1 isolate LAV (HIV-1) (K02013) from Wain-Hobson et al. (1985); human T-cell leukemia/lymphoma virus type I (HTLV-I) (J02029, M33896) from Seiki et al. (1983); mouse mammary tumor virus (MMTV) (M15122) from Moore et al. (1987); Rous sarcoma virus (RSV) (J02021, J02342, J02343) from Schwartz et al. (1983); intracisternal type A particle (IAP) (M23189) from Mietz et al. (1987); human spuma retrovirus (HSRV) (M19427) from Maurer et al. (1988); Moloney murine leukemia virus (MoMLV) (M76668) from Shinnick et al. (1981); Mason–Pfizer monkey virus (MPMV) (M12349) from Sonigo et al. (1986); and simian SRV-1 type D retrovirus (SRV-1) (M11841) from Power et al. (1986).

Results

Biased Nucleotide Composition of HERV-K Genomes

The 34 HERV-K sequences analyzed in this study represent four HML groups and are listed in Table 1. First,

Table 1. Nucleotide composition of HERV-K RNA genomes[a]

	A (%)	C (%)	G (%)	T (%)	CG/GC	Ratio
HERV-K isolates						
HML-1 group ($n = 1$)						
St.4[b]	37.5	18.3	16.5	27.7	5/28	0.18
HML-2 group ($n = 21$)						
L4.4	34.3	18.5	18.2	29.1	5/27	0.19
H3.1	34.3	18.4	18.2	29.1	5/27	0.19
P1.1[b]	34.3	19.0	18.0	28.7	8/27	0.30
HERVK10[b]	34.6	19.0	18.0	28.4	8/26	0.31
N8.4[b]	34.5	18.8	18.0	28.7	8/26	0.31
HERVK55	34.2	18.9	18.2	28.8	8/26	0.31
HERVK27	34.7	19.0	18.0	28.3	9/25	0.36
HERVK67[b]	34.1	19.3	18.3	28.2	8/27	0.30
T47D[b]	34.5	18.7	18.0	28.9	6/27	0.20
P1.8[b]	33.9	19.2	18.5	28.3	9/29	0.31
M3.5[b]	34.1	19.0	18.4	28.5	8/29	0.28
St.2[b]	33.9	19.0	18.3	28.7	10/29	0.34
M3.8	33.8	20.0	18.3	28.0	8/27	0.30
HP.1	33.8	19.8	18.1	28.3	9/26	0.35
P1.4	34.4	19.7	17.9	28.0	13/24	0.54
P1.10	34.2	19.8	18.1	27.9	13/25	0.52
D1.3	33.6	19.1	17.3	30.0	5/25	0.20
HM16	33.9	19.0	17.1	30.0	4/25	0.16
D1.2	33.5	19.4	18.7	28.5	4/28	0.14
K1.1	33.7	19.4	18.4	28.6	3/28	0.11
HERVK73	33.6	19.5	18.5	28.4	5/28	0.18
mean	34.1	19.2	18.1	28.6		
SD	±0.3	±0.4	±0.4	±0.6		
HML-3 group ($n = 9$)						
HERV76	31.4	19.8	16.6	32.1	5/22	0.23
M3.9	31.5	20.6	17.3	30.5	7/26	0.27
P1.3	33.7	19.7	16.4	30.2	4/27	0.15
N5.1	32.4	19.7	16.5	31.4	4/25	0.08
HERV50	33.5	20.9	15.8	29.8	7/26	0.27
HP.2	32.4	20.1	17.0	30.5	6/23	0.26
M3.1	33.2	19.9	15.9	31.0	2/24	0.08
P1.6	33.8	20.2	16.0	30.1	1/26	0.03
P1.7	33.8	20.2	16.0	30.1	1/26	0.04
mean	32.9	20.1	16.4	30.6		
SD	±0.9	±0.4	±0.5	±0.7		
HML-4 group ($n = 3$)						
St.1[b]	28.4	22.9	16.6	32.1	5/29	0.17
ERVMLN	27.0	24.5	17.8	30.8	7/29	0.24
M3.10[b]	28.2	22.9	17.0	31.9	7/27	0.26
mean	27.9	23.4	17.1	31.6		
SD	±0.8	±0.9	±0.6	±0.7		
Other retroviruses						
IAP	30.1	22.0	18.8	29.1	2/19	0.11
HIV-1	38.3	17.3	20.8	23.6	1/15	0.07
HTLV-I	25.8	34.6	16.7	22.8	5/37	0.41
MMTV	35.3	18.3	17.6	28.7	5/19	0.26
RSV	20.6	25.8	27.4	26.2	26/44	0.59
HSRV	36.9	15.7	16.4	31.0	2/16	0.15
MPMV	34.3	20.2	16.5	29.1	5/19	0.26
MoMLV	26.4	28.7	25.0	19.9	15/29	0.52
SRV-1	34.3	20.3	16.6	29.1	9/24	0.38

[a] All values listed concern the genomic (+) RNA strand.
[b] HML RT segments with an open reading frame.

we compared the nucleotide composition of the (+) strand RT segment with that of the equivalent RT fragment of a set of exogenous retroviruses. Like most retroviral species, the HERV-K elements were found to have an asymmetric and biased genome composition. A preference for the T nucleotide, and in particular for the A nucleotide, can easily be recognized in these sequences (Table 1). The average A count ranges from 32.9% for HML-3 to 37.5% for the single member of the HML-1 group (clone St.4). Although the HML-4 group maintains a preference for the A nucleotide (27.9%), this group is rather special in having relatively high T and C counts (31.6 and 23.4%, respectively). In fact, HML-4 represents the first retroviral species with T as the most abundant nucleotide in the viral genomic (+) strand. Among the retroviridae, the HML-3 group, together with the exogenous human spuma retrovirus (HSRV), sets another record, with a minimal G count of only 16.4%.

The nucleotide content of the retroviral genomes is depicted in Fig. 1. The sequences are ordered according to their A count, starting with RSV (20.6% A) and ending with HIV-1 (38.3%). This survey indicates that the base count of the HERV-K elements is dramatically different from that of several other retroviruses such as RSV (A-poor), HTLV-I (C-rich, G-poor), and MoMLV (the only retrovirus with a relatively unbiased genome composition). On the other hand, characteristics of the HERV-K family (A- and T-rich) are shared by several retroviral species (SRV-1, MPMV, MMTV, HSRV, HIV-1) and are the intracisternal type A particles (IAP). To reveal potential rules for genomes with such a biased nucleotide composition, we made a pairwise comparison of the base counts. The variation of the C, T, and G counts as a function of the A count is plotted in Fig. 2 for this set of retroviruses. Interestingly, comparison of the A and C counts revealed that an increase in the number of A nucleotides correlates with a reduction in the number of C nucleotides. The notable exception is RSV, which is the only retrovirus with a relatively unbiased genome composition, approximately 25% of each nucleotide. No correlations are evident for other nucleotides (Fig. 2 and data not shown).

It can be argued that the particular nucleotide composition of the HERV-K genome is restricted to the 600-nt RT fragment that was analyzed. We therefore extended this analysis for the full-length genome of the prototype HERV-K10 isolate, a member of the HML-2 subgroup (Ono et al. 1986). The graph in Fig. 3 clearly demonstrates that the A nucleotide is favored over the complete genome length, with the possible exception of the extreme 5′ and 3′ ends, where the A count drops to approximately 25%. A similar trend was described for HIV-1, in which case it has been argued that the presence of important regulatory signals in the long terminal repeat elements may contribute to this effect (van Hemert

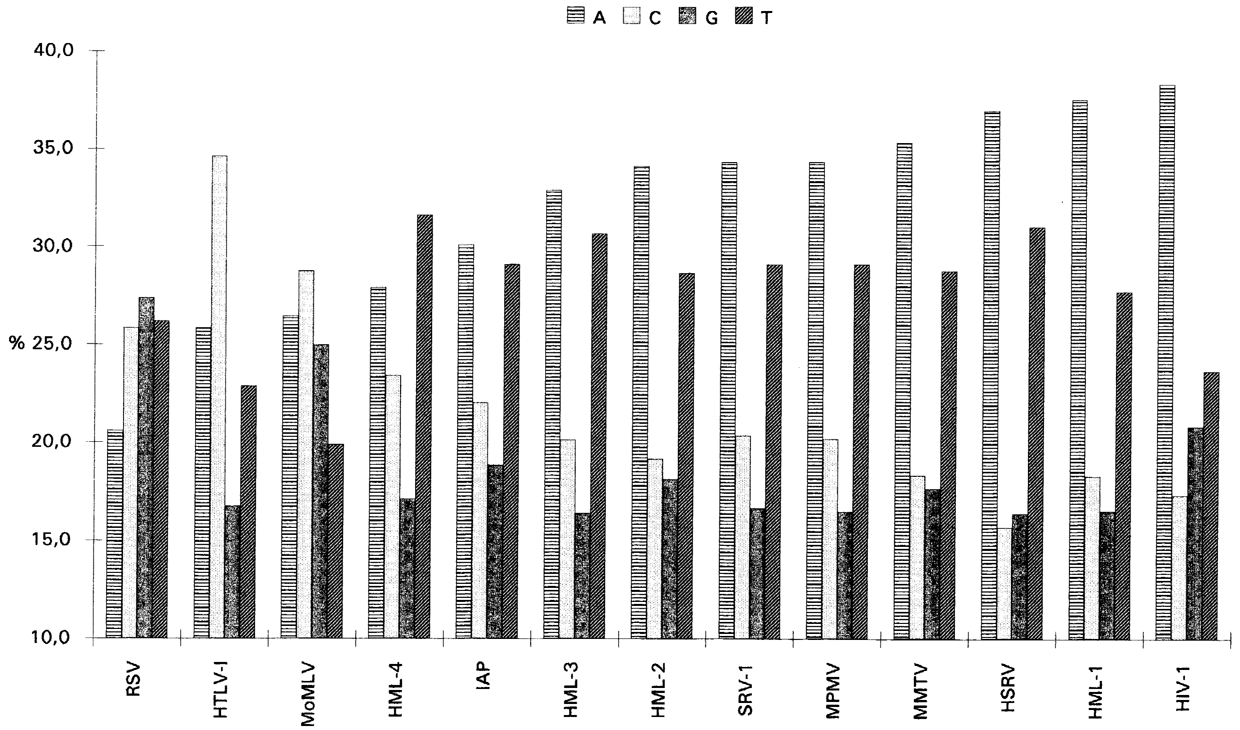

Fig. 1. Nucleotide composition (percentage A, C, G, T) of retroviral genomes. The values represent the relative frequency of the four nucleotides in the genomic (+) strand. The analysis was restricted to the 600-nt RT segment. We analyzed both a set of exogenous retroviruses and the four HML groups of the endogenous HERV-K viruses. The viruses were ranked according to their A count, from low (RSV, 20.6%) to high (HIV-1, 38.3%). The values for individual members of the HML groups are listed in Table 1.

and Berkhout 1995). Specifically, recognition of the A-rich transcriptional initiation and termination signals (TATAA box and AATAAA motif, respectively) may be enhanced in an environment that does not have a surplus of the A nucleotide. Two additional remarks can be made. First, although some local fluctuation can be observed, all four base counts are relatively stable over the whole genome length. Second, careful examination of the graphs in Fig. 3 further strengthens the idea that the A and C counts are inversely correlated. For instance, peak values of the A content around genome position 1100–1700 are compensated for by a decrease in the C count, whereas the relative frequency of the other nucleotides remains largely unchanged.

Surprinsingly, the preference of the HERV-K genome for A nucleotides is not more enhanced at the synonymous codon positions (Fig. 4). The A nucleotides are distributed almost equally among the three codon positions and are the dominant base at all positions, except for the HML-4 subgroup. In contrast, a strong effect of codon position is apparent for C, which is highly suppressed at the third codon position. Codon position effects are also obvious for G and T. The G suppression is most evident at the second codon position, whereas nearly normal levels are reached at the first codon position. Another pattern is observed for the preferred T nucleotide, which is favored only at the second and third codon positions.

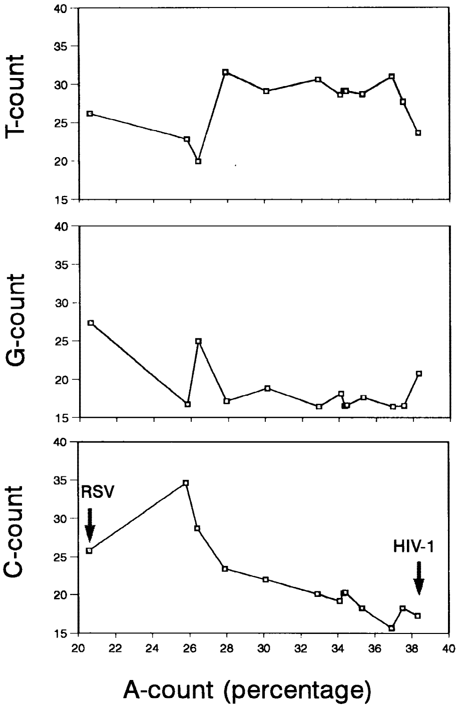

Fig. 2. Inverse correlation between the A and the C content of retroviral genomes. For the retroviral species listed in Fig. 1, we plotted the T, G, and C counts of the viral (+) strand as a function of the A count (plotted on the *x*-axis). The ranking order of the retroviruses is the same as in Fig. 1, going from A-poor genome (RSV, 20.6%) to A-rich genomes (HIV-1, 38.3%). Other nucleotide relations were analyzed (not shown) but did not show a correlation.

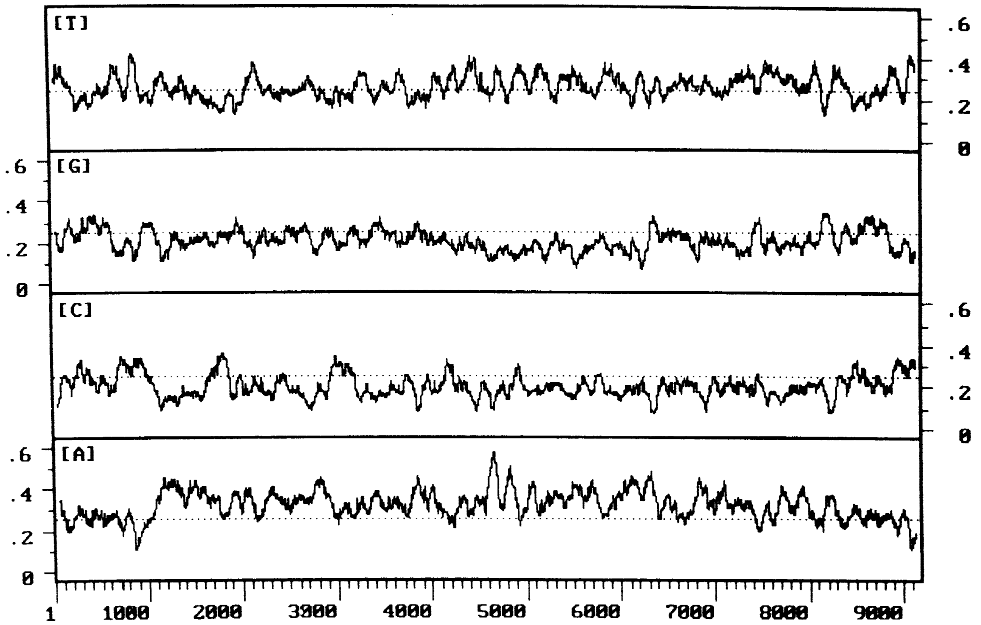

Fig. 3. Nucleotide sequence composition along the complete HERV-K10 genome. The nucleotide content of the full-length genome (9179 nucleotides) was analyzed with the PBASE program of the PC/Gene package (window setting, 100 nucleotides). The *dotted lines* represent the 0.25 value, which is the expected nucleotide ratio for a random sequence.

CpG Suppression in HERV-K Genomes

We analyzed the genome of the endogenous HERV-K retrovirus for possible CpG suppression. A standard analysis of the dinucleotide count is hampered by the fact that the relative frequency of both the C and the G nucleotides is very low in these genomes (see Table 1). Thus, for calculation of the CpG frequency one should correct for this bias or, alternatively, compare the frequency of the CpG dinucleotide with that of the GpC motif. This ratio will be about 1.0 in a gene fragment with a random nucleotide sequence. Results of the CpG analysis of all HERV-K members and the set of exogenous retroviruses are listed in Table 1 and presented graphically in Fig. 5. CpG suppression was evident for the RT fragment of all four HML groups. The effect appears to be most pronounced for the HML-3 group and least spectacular for the HML-2 group. This bias against CpG is a property that extends over the whole HERV-K genome length, as verified for the HERV-K10 isolate. The full-length HERV-K10 genome has a CpG/GpC ratio of 0.36, which is very similar to the 0.31 value of the RT fragment (Table 1). Discrimination against CpG appears to be very specific, because an extensive analysis of the HERV-K sequences did not reveal any other dinucleotide sequence motifs that are either favored or disfavored that strongly. For other dinucleotides, the bias ranged from 0.78 (TA dinucleotide) to 1.37 (CC dinucleotide). A wide variety in CpG suppression levels was apparent for all other viruses (Table 1), with CpG/GpC ratios ranging from 0.07 [HIV-1 (see also Kypr et al. 1989; Shpaer and Mullins 1990)] to 0.59 (RSV). Thus, the endogenous HERV-K retrovirus resembles most exogenous counterparts in having a strong bias against the CpG dinucleotide.

Selective Pressure to Conserve the HERV-K RT Open Reading Frame

The maintenance of open reading frames, in particular that of the RT gene, in a significant number of HERV-K elements has raised the possibility that these elements are still active in retrotransposition. Evaluation of synonymous versus nonsynonymous substitutions can provide additional insight into the evolutionary pressures acting on a nucleic acid sequence. The synonymous distance (d_s), and nonsynonymous distance (d_a) were calculated for the 600-nt RT fragment both within and between HML groups (Table 2). To calculate the mean synonymous and nonsynonymous distances between two separate HML groups, each sequence belonging to one group was compared with each sequence from the other group. The ratio of these values (d_s/d_a) is listed in Table 2. We found that most sequence variation in the HERV-K sequences resulted from synonymous nucleotide substitutions, pointed out by the high (d_s/d_a) ratios. The values of synonymous substitutions between the groups were extremely high and reached the level of saturation. These results indicate that purifying selection is currently operating on the endogenous RT elements or was operating on the RT gene as part of the corresponding exogenous viruses.

Mutational Bias of the HERV-K RT Enzyme

The typical nucleotide composition of HERV-K genomes may be caused by characteristic misincorporation

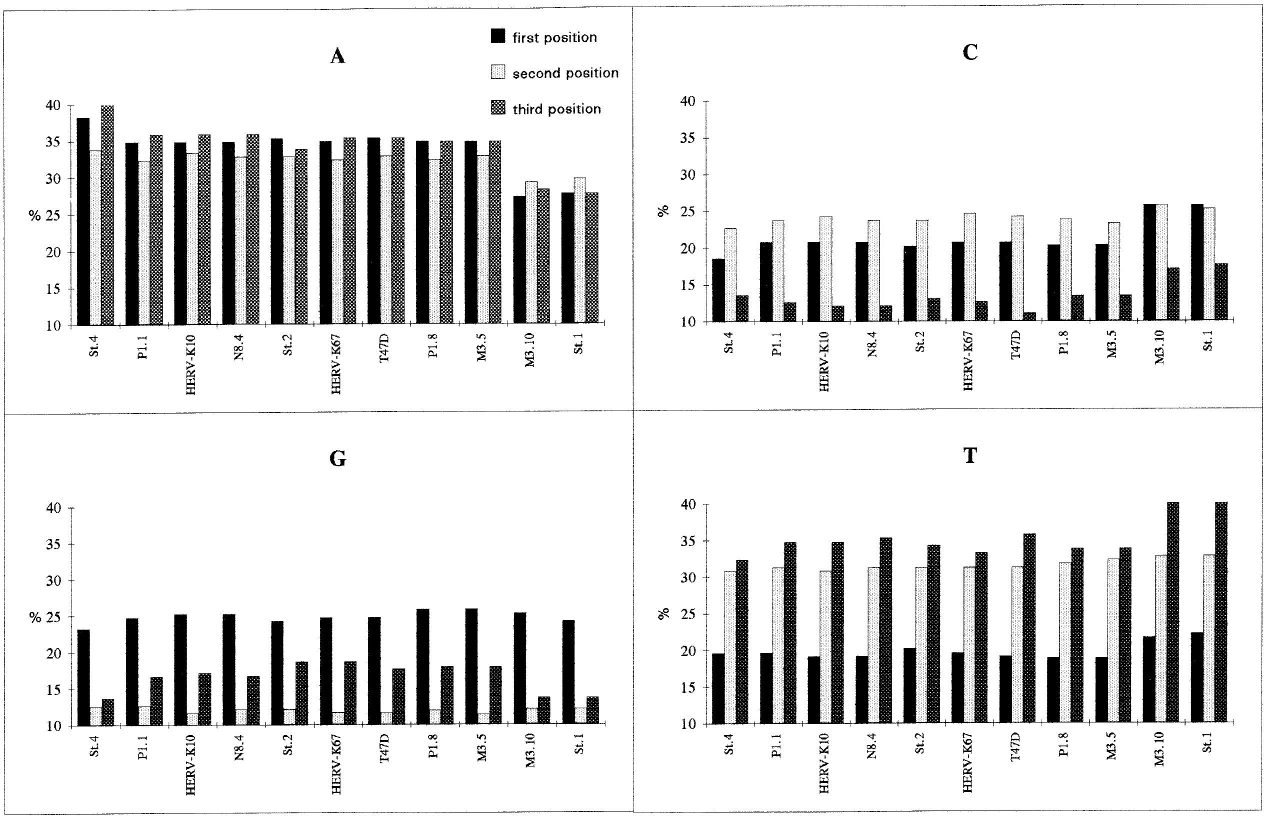

Fig. 4. Nucleotide composition at different codon positions. HERV-K elements with an open RT reading frame were selected for this analysis. Their overall nucleotide count is depicted in Fig. 1 and listed in Table 1. Here we analyzed the distribution of the four nucleotides at the first, second, and third codon positions. The HERV-K isolates are ranked according to their A count, from high (St.4, 37.5%) to low (St.1, 28.4%).

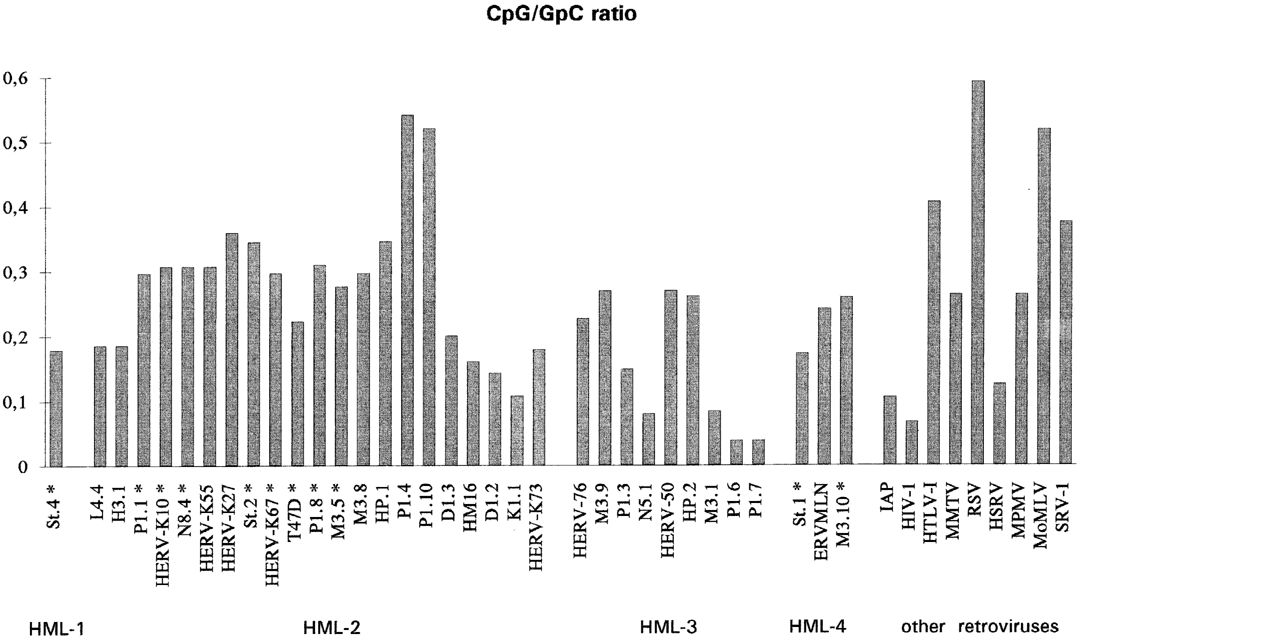

Fig. 5. CpG methylation sites are underrepresented in HERV-K genomes. The total number of CpG and GpC dinucleotides within the approximately 600-nt RT fragement [(+) strand only] are listed in Table 1. The CpG/GpC ratio for a random sequence will be about 1.0. HERV-K isolates with an open RT reading frame are marked with an *asterisk.*

Table 2. Synonymous and nonsynonymous distances between HERV-K RT genes

HML group	Synonymous distance (d_s)	Nonsynonymous distance (d_a)	d_s/d_a
1 (n = 1)	x[a]	x	x
2 (n = 21)	0.157	0.036	4.33
3 (n = 9)	0.170	0.096	1.77
4 (n = 3)	0.394	0.090	4.38
1 vs 2	0.670	0.156	>4.30[b]
1 vs 3	0.750	0.208	>3.61[b]
1 vs 4	0.724	0.192	>3.77[b]
2 vs 3	0.756	0.221	>3.42[b]
2 vs 4	0.763	0.197	>3.87[b]
3 vs 4	0.814	0.257	>3.16[b]

[a] Not done (only one sequence in the group).

[b] As synonymous substitutions between the groups reached the level of saturation, d_s/d_a data represent minimum values.

properties of the viral RT enzyme. For instance, it has been suggested that the ability of the HIV-1 RT enzyme to extend efficiently G.dT mismatches results in frequent G-to-A changes and accumulation of the A nucleotide (see Discussion). We therefore wanted to analyze the pattern of nucleotide substitutions for the HERV-K family. Because data on de novo mutations are not available for this endogenous retrovirus, we relied on nucleotide sequence comparisons within the different HERV-K groups with more than one member (HML-2, HML-3, and HML-4). We estimated the pattern of nucleotide substitutions as follows. To infer the direction of substitution, we took the consensus sequence of each HML group as the precursor sequence and scored the substitutions seen in individual members of the HML groups. Each mutation was logged only once to avoid the multiple inclusion of base changes that are identical by descent. Obviously, one cannot formally exclude the possibility that these sites actually do represent truly hot spots of mutation and/or selection.

Nucleotide substitution rates, obtained for the 600-nt RT fragment of the three HML groups, are depicted in Fig. 6. As a control, we included the mutational pattern of another A-rich retrovirus, the exogenously replicating HIV-1. The frequency of transitional changes was high for all HML groups compared with the frequency of transversions. Among the transitions, a significant preference for the C-to-T substitution was observed, consistent with the low C count and high T count of these viruses. The HML-2 group, which has the highest A count of the three groups analyzed (34.1%; see Table 1), demonstrated a dominance of the G-to-A mutation. The latter mutational pattern is very similar to that of HIV-1 (Moriyama et al. 1991), for which G-to-A mutation has been implicated in the evolution toward A-rich genomes (see Discussion).

Discussion

All actively replicating exogenous retroviruses have a typical genome composition that is both biased with respect to usage of the four nucleotides and asymmetric regarding the two genome strands (Bronson and Anderson 1994; Berkhout and van Hemert 1994; Zoubak et al. 1992; Kypr et al. 1989). This study indicates that these characteristics, as well as several other peculiarities of retroviral genomes, are not unique properties of exogenously replicating retroviruses, as some of them are also observed for the HERV-K family of endogenous retroviral elements. The HERV-K elements are particularly A-rich and, to a lesser extent, also T-rich. An overview of a large set of retrovirus genomes indicates that the extent of accumulation of the A nucleotide is inversely correlated with the C content of the genome, which may represent a compensation mechanism. Because it is generally assumed that the HERV-K elements have undergone no or only a limited number of replication cycles following their entry into the human germline some 30 million years ago, it is likely that the hypothetical, exogenous precursor of these HERV-K elements are similarly A/T-rich, suggesting that biased genome composition is an ancient property of the retroviridae. These integrated HERV-K proviruses will also have been subject to spontaneous mutation as part of the host chromosome. This process will account for approximately 6% genome variation for elements that integrated 30 million years ago (Li and Grau 1991). By sequence comparison of the different HERV-K members, we present preliminary evidence for a G-to-A and C-to-T mutational bias of the HERV-K RT enzyme. Thus, in analogy with the HIV-1 system, this may indicate that the retroviral RT enzyme was directly involved in shaping of the HERV-K genome. Identification of an enzymatically active HERV-K RT enzyme will allow further testing of this hypothesis.

Although the molecular mechanism(s) responsible for the drift of retroviral genomes toward a particular nucleotide composition is (are) not known, two possibilities have been suggested (Berkhout and van Hemert 1994). Selection for particular features in either the viral RNA genome or the proviral DNA genome may have been the driving force in the generation of biased genomes. For instance, retroviruses have a pattern of codon usage that is different from that of their host cells, although there is currently no evidence that this feature provides the virus with a regulatory mechanism (van Hemert and Berkhout 1995; Haas et al. 1996). Alternatively, the typical genome composition may be the result of biased nucleotide incorporation properties of the viral RT polymerase. There is accumulating evidence for this mutation-driven scenario in the case of the HIV-1 retrovirus, for which it has been reported that the nucleotide misincorporation spectrum of its RT enzyme is responsible for the A-richness of the HIV-1 genome. In par-

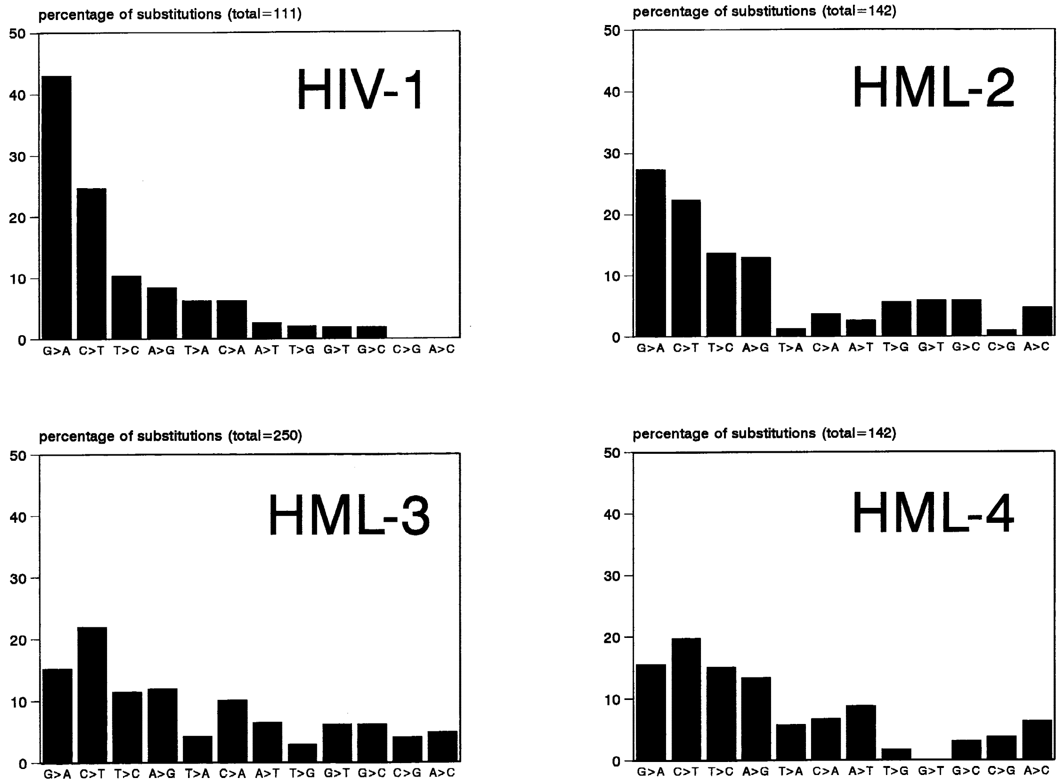

Fig. 6. Mutational spectrum of endogenous HERV-K elements and the exogenous HIV-1 virus. The HIV-1 data were derived from several long-term tissue culture replication studies with mutant viruses (Klaver and Berkhout 1994; Berkhout et al. 1997). A total of 111 nucleotide substitutions was scored in the HIV-1 leader region of TAR- and poly(A)-hairpin mutants. Since these mutants do not revert to a particular sequence (e.g., the wild-type sequence), no bias was introduced in the analysis. Analysis of the HERV-K mutational spectrum was performed as described under Materials and Methods. The total number of nucleotide substitutions analyzed is listed (e.g., 250 for the HML-3 group). The 12 possible mutations (4 transitions, 8 transversions) were ranked based on the HIV-1 system, from high (G-to-A) to low (A-to-C).

ticular, it has been suggested that a skewed dCTP pool, combined with the stability of G–dT mispairs, eventually leads to the accumulation of A's in the viral genome (Vartanian et al. 1994). Strand specificity can be explained since such misincorporations became fixated during first-strand synthesis (RNA-dependent DNA polymerization) because RNaseH will degrade the RNA template. However, similar errors during second-strand synthesis (DNA-dependent DNA polymerization) will be subjected to the DNA repair machinery of the cell upon integration of the proviral DNA. This mechanism will lead to a gradual drift of the genomic RNA (+) strand toward A abundance. Obviously, this trend may eventually restrain the fitness of the viral progeny. Perhaps to cope with such an evolutionary restriction, several A-rich retroviruses such as caprine arthritis–encephalitis virus (CAEV) encode a dUTPase activity that can limit this drift (Turelli et al. 1997).

As most nucleotide substitutions at the third position of a coding triplet are synonymous, these positions are relatively free from functional constraints, at least at the protein level. For this reason, a drift toward a particular nucleotide composition can be especially pronounced at such ''silent'' codon positions. This phenomenon was indeed observed for the A-rich genome of HIV-1. This exogenous retrovirus has an average A count of 39.0% but reaches an extreme value of 46.5% at the silent codon position (Berkhout and van Hemert 1994; van Hemert and Berkhout 1995). Interestingly, no such effect was apparent for the abundant A nucleotide of the HERV-K genome, but such a trend was seen for the restricted C nucleotide. The T and G nucleotides also show a particular distribution over the three codon positions. The preference for G nucleotides at the first codon position may be related to the infrequent use of C nucleotides at the third codon position, thereby precluding the formation of CpG methylation sites (see below). The underlying causes of other codon positional effects, in particular, the even distribution of the A nucleotide, remain unclear.

Differences in the base count and RT mutational spectrum were observed for the separate HML groups. Most notably, the two most abundant genomic nucleotides varied among the HML groups (e.g., HML-1, 37.5% A and 27.7% T; HML-4, 27.9% A and 31.6% T). The G-to-A mutational bias was most evident for the HML-2 group, which also has the most extreme A count among the three HML groups for which the mutational spectrum could be estimated. These different genome characteristics fully support the HML phylogeny proposed for the HERV-K family (Medstrand and Blomberg 1993). Assuming that only a limited number of intracellular HERV-K replication cycles has taken place, it is likely that these contemporary HML groups represent different clades of an exogenous HERV-K family that became fixated in the genome by independent viral introductions.

As shown in this study, ancient retroviral sequences display significant suppression of CpG methylation sites. Paucity of the CpG dinucleotide sequence is common to all small eukaryotic DNA and RNA viruses, including the current exogenous retroviruses (Karlin et al. 1994). Several mechanistic explanations have been proposed. Selection against CpG sites may be driven by their property to be hypermutable upon methylation. 5-Methylcytosine (5mC) occurs predominantly in CpG dinucleotides and is prone to deamination to form thymidine (T). Alternatively, there is some evidence that methylation can interfere with transcription of viral genes (Shpaer and Mullins 1990). Suppression of the CpG methylation sites may enhance efficient transcription of the integrated provirus.

Several observations suggest that selective pressure continued to be operating on at least some of the HERV-K elements after they integrated into the human genome. First, analysis of the synonymous and nonsynonymous mutation rates revealed a significant bias for the synonymous or silent codon changes (d_s/d_a ratios ranged from 1.77 to >4.38). Similar ratios (>1) can be measured for the RT gene of actively replicating retroviruses such as HIV-1 (Cornelissen et al. 1997), while in the HIV-1 regions which are under strong immune pressure, such as the envelope V3 region, d_s/d_a ratios below 1 have been reported (Lukashov et al. 1995). The strong prevalence of the synonymous nucleotide substitutions indicates that purifying (negative) selection is or has been operating on the HERV-K family and correlates with the maintenance of open reading frames in many HERV-K members, in particular, those of the HML-2 group. The d_s/d_a ratios may suggest that the contemporary, endogenized HERV-K elements are under continuous selective pressure to maintain replication/retrotransposition capacity. The latter possibility is consistent with the notion that some of the HML groups contain a very large number of members. For instance, at least 23 members of the HML-2 group have currently been identified, but there is evidence for the presence of many additional copies within the human genome (Zsíros and Berkhout, unpublished results). Another line of evidence that is consistent with the idea that HERV-K elements have remained active is the demonstration of RT enzyme activity within retroviral particles (Simpson et al. 1996; Patience et al. 1996; Conrad et al. 1997). It will be important to identify and analyse these active HERV-K members in more detail. Such studies should also shed more light on the possibility that these elements are involved in the evolution of the human genome by retrotransposition events.

Finally, we mentioned previously that the biased genome composition of retroviral genomes could bias phylogenetic analyses that are based solely on similarity in nucleotide sequence (Berkhout and van Hemert 1994). Specifically, it was argued that similarity in sequence between two virus groups may not necessarily indicate a close evolutionary relationship but, instead, may reflect convergent evolution driven by unrelated RT enzymes with comparable misinsertion properties. The problem is particularly evident for the many retroviruses that share the HERV-K-like genome composition (A- and T-rich, G- and C-poor). This group includes many exogenous species (SRV-1, MPMV, MMTV, HSRV, HIV-1; see Fig. 1). Biased genome composition should be taken into account when the mutation rates and evolutionary distances of these exogenous and endogenous retroviral species are calculated.

Acknowledgments. We thank Formijn van Hemert for critical reading of the manuscript. This research was supported by the ‘‘Stichting Kindergeneeskundig Kankeronderzoek’’ (SKK).

References

Andersson M-L, Medstrand P, Yin H, Blomberg J (1996) Differential expression of human endogenous retroviral sequences similar to mouse mammary virus in normal peripheral blood mononuclear cells. AIDS Res Hum Retrovirus 12:833–840

Berkhout B, van Hemert FJ (1994) The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. Nucleic Acids Res 22:1705–1711

Berkhout B, Klaver B, Das AT (1997) Forced evolution of a regulatory RNA helix in the HIV-1 genome. Nucleic Acids Res 25:940–947

Bronson EC, Anderson JN (1994) Nucleotide composition as a driving force in the evolution of retroviruses. J Mol Evol 38:506–532

Conrad B, Weissmahr RN, Boni J, Arcari R, Schupbach J, Mach B (1997) A human endogenous retroviral superantigen as candidate autoimmune gene in type 1 diabetes. Cell 90:303–313

Cornelissen M, van den Burg R, Zorgdrager F, Lukashov V, Goudsmit J (1997) Pol gene diversity of five human immunodeficiency virus type 1 subtypes: Evidence for naturally occurring mutations that contribute to drug resistance, limited recombination patterns, and common ancestry for subtypes B and D. J Virol 71:6348–6358

Deen KC, Sweet RW (1986) Murine mammary tumor virus pol-related sequences in human DNA: Characterization and sequence comparison with the complete murine mammary tumor virus pol gene. J Virol 57:422–432

Haas J, Park E-C, Seed B (1996) Codon usage limitation in the expression of HIV-1 envelope glycoprotein. Curr Biol 6:315–324

Karlin S, Doerfler W, Cardon LR (1994) Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J Virol 68:2889–2897

Klaver B, Berkhout B (1994) Evolution of a disrupted TAR RNA hairpin structure in the HIV-1 virus. EMBO J 13:2650–2659

Kypr J, Mrazek J, Reich J (1989) Nucleotide composition bias and CpG dinucleotide content in the genomes of HIV and HTLV 1/2. Biochim Biophys Acta 1009:280–282

Li MD, Bronson DL, Lemke TD, Faras AJ (1995) Restricted expression of new HERV-K members in human teratocarcinoma cells. Virology 208:733–741

Li MD, Lemke TD, Bronson DL, Faras AJ (1996) Synthesis and analysis of a 640-bp pol region of novel human endogenous retroviral sequences and their evolutionary relationships. Virology 217:1–10

Li W, Graur D (1991) Fundamentals of molecular evolution. Sinauer Associates, Sunderland, MA

Lower R, Boller K, Hasenmaier B, Korbmacher C, Mueller-Lantzsch N, Lower J, Kurth R (1993) Identification of human endogenous retroviruses with complex mRNA expression and particle formation. Proc Natl Acad Sci USA 90:4480

Lower R, Lower J, Kurth R (1996) The viruses in all of us: Characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci USA 93:5177–5184

Lukashov VV, Kuiken CL, Goudsmit J (1995) Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period. J Virol 69:6911–6916

Maurer B, Bannert H, Darai G, Glugel RM (1988) Analysis of the primary structure of the long terminal repeat and the gag and pol genes of the human spumaretrovirus. J Virol 62:1590–1597

Medstrand P, Blomberg J (1993) 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 67:6778–6787

Medstrand P, Linderskog M, Blomberg J (1992) Expression of human endogenous retroviral sequences in peripheral blood mononuclear cells of healthy individuals. J Gen Virol 73:2463–2466

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

Mietz JA, Grossman Z, Lueders KK, Kuff EL (1987) Nucleotide sequence of a complete mouse intracisternal A-particle genome: Relationship to known aspects of particle assembly and function. J Virol 61:3020–3029

Moore R, Dixon M, Smith RE, Peters G, Dickson C (1987) Complete nucleotide sequence of a milk-transmitted mouse mammary tumor virus: Two frameshift suppression events are required for translation of gag and pol. J Virol 61:480–490

Moriyama EN, Ina Y, Ikeo K, Shimizu N, Gojobori T (1991) Mutation pattern of human immunodeficiency virus genes. J Mol Evol 32: 360–363

Mueller-Lantzsch N, Sauter M, Weiskircher A, Kramer K, Best B, Buck M, Grasser F (1993) Human endogenous retroviral element K10 (HERV-K10) encodes a full-length gag homologous 73-kDa protein and a functional protease. AIDS Res Hum Retrovirus 9: 345–350

Ono M, Yasunaga T, Miyata T, Ushikubo H (1986) Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. J Virol 60:589–598

Patience C, Simpson GR, Colletta AA, Welch HM, Weiss RA, Boyd MT (1996) Human endogenous retrovirus expression and reverse transcriptase activity in the T47D mammary carcinoma cell line. J Virol 70:2654–2657

Power MD, Marx PA, Bryant ML, Gardner MB, Barr PJ, Luciw PA (1986) Nucleotide sequence of SRV-1, a type D simian acquired immuno deficiency syndrome retrovirus. Science 231:1567–1572

Schwartz De, Tizard R, Gilbert W (1983) Nucleotide sequence of Rous sarcoma virus. Cell 32:853–869

Seifarth W, Skladny H, Krieg-Schneider F, Reichert A, Hehlmann R, Leib-Mosch C (1995) Retrovirus-like particles released from the human breast cancer cell line T47-D display type B- and C-related endogenous retroviral sequences. J Virol 69:6408–6416

Seiki M, Hattori S, Hirayama Y, Yoshida M (1983) Human adult T-cell leukemia virus: Complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. Proc Natl Acad Sci USA 80:3618–3622

Shinnick TM, Lerner RA, Sutcliffe JG (1981) Nucleotide sequence of Moloney murine leukaemia virus. Nature 293:543–548

Shpaer EG, Mullins JI (1990) Selection against CpG dinucleotides in lentiviral genes: A possible role of methylation in regulation of viral expression. Nucleic Acids Res 18:5793–5797

Simpson GR, Patience C, Lower R, Tonjes RR, Moore HDM, Weiss Ra, Boyd MT (1996) Endogenous D-type (HERV-K) related sequences are packaged into retroviral particles in the placenta and possess open reading frames for reverse transcriptase. Virology 222:451–456

Sonigo P, Barker CS, Hunter E, Wain-Hobson S (1986) Nucleotide sequence of Mason-Pfizer monkey virus: An immunosuppressive D-type retrovirus. Cell 45:375–385

Steinhuber S, Brack M, Hunsmann G, Schwelberger H, Dierich MP, Vogetseder W (1995) Distribution of human endogenous retrovirus HERV-K genomes in humans and different primates. Hum Genet 96:188–192

Turelli P, Guiguen F, Mornex J-F, Vigne R, Querat G (1997) dUTPase-minus caprine arthritis-encephalitis virus is attenuated for pathogenesis and accumulates G-to-A substitutions. J Virol 71:4522–4530

van Hemert FJ, Berkhout B (1995) The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. J Mol Evol 41:132–140

Vartanian J-P, Meyerhans A, Sala M, Wain-Hobson S (1994) G $\geqslant$ A hypermutation of the human immunodeficiency virus type 1 genome: Evidence for dCTP pool imbalance during reverse transcription. Proc Natl Acad Sci USA 91:3092–3096

Wain-Hobson S, Sonigo P, Danos O, Cole S, Alizon M (1985) Nucleotide sequence of the AIDS virus, LAV. Cell 40:9–17

Wilkinson DA, Mager DL, Leong J-A (1994) Endogenous human retroviruses. In: Levy JA (ed) The Retroviridae. Plenum Press, New York, pp 465–535

Zoubak S, Rynditch A, Bernardi G (1992) Compositional bimodality and evolution of retroviral genomes. Gene 119:207–213

Zsíros J, Jebbink MF, Lukashov VV, Voûte PA, Berkhout B (1998a) Evolutionary relationships within a subgroup of HERV-K-related human endogenous retroviruses. J Gen Virol 79:61–70

Zsíros J, Jebbink MF, Voûte PA, Berkhout B (1998b) Identification of novel human endogenous retroviral sequences belonging to the HERV-K family. AIDS Res Hum Retrovir 12:1093–1098.