ELSEVIER

Gene 273 (2001) 51–61

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Full-sized HERV-K (HML-2) human endogenous retroviral LTR sequences on human chromosome 21: map locations and evolutionary history

Sergey G. Kurdyukov[a], Yuri B. Lebedev[a,*], Irena I. Artamonova[a], Tatyana N. Gorodentseva[a], Anastasia V. Batrak[a], Ilgar Z. Mamedov[a], Tatyana L. Azhikina[a], Svetlana P. Legchilina[b], Irina G. Efimenko[b], Katheleen Gardiner[c], Eugene D. Sverdlov[a,b]

[a]*Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Science, 16/10 Miklukho-Maklaya, Moscow, 117871, Russia*
[b]*Institute of Molecular Genetics, Russian Academy of Science, Moscow, Russia*
[c]*Eleanor Roosevelt Institute, Denver, CO, USA*

Received 26 March 2001; received in revised form 11 May 2001; accepted 14 June 2001

Abstract

One of the evolutionary mechanisms for acquisition of novel functional sequences can be domestication of exogenous retroviruses that have been integrated into the germ line. The whole genome mapping of such elements in various species could reveal differences in positions of the retroviral integration and suggest possible roles of these differences in speciation. Here, we describe the number, locations and sequence features of the human endogenous retrovirus HERV-K (HML-2) long terminal repeat (LTR) sequences on human chromosome 21. We show that their distribution along the chromosome is not only non-random but also roughly correlated with the gene density. Amplification of orthologous LTR sites from a number of primate genomes produced patterns of presence and absence for each LTR sequence and allowed determination of the phylogenetic ages and evolutionary order of appearance of individual LTRs. The identity level and phylogenetic age of the LTRs did not correlate with their map locations. Thus, despite the non-random distribution of LTRs, they have apparently been inserted randomly into the chromosome relative to each other. As evidenced in previous studies of chromosomes 19 and 22, this is a characteristic of HERV-K integration.

Keywords: Human endogenous retrovirus-K; HML-2; Long terminal repeat; Primate evolution; Chromosome 21; Human genome

1. Introduction

The human genome contains above 40% of DNA sequences derived from different types of transposable elements (Smit, 1999). In particular, recently sequenced human chromosomes 21 (Hattori et al., 2000) and 22 (Dunham et al., 1999) comprise up to 40 and 42%, respectively, of these interspersed repeats at one time actively transposed across the genome. But then the transposon activity apparently ceased, although some of the elements (mainly Alu and L1) are still being transposed from time to time disclosing themselves through deleterious mutations that cause hereditary diseases (reviewed in Smit, 1999; Kazazian, 2000). Being a large portion of the genome, the transposons should unavoidably have played an important role in the evolution, in particular providing new genes as well as new regulatory elements for multitudes of functional genes (reviewed in Britten, 1997; Smit, 1999; Kazazian, 2000; Brosius, 1999). However, despite a good number of examples of the involvement of former transposons in the particular gene functions (reviewed in Britten, 1997; Smit, 1999; for recent results see Hamdi et al., 2000), clear understanding of the evolutionary impact(s) of a given transposition in a given genomic site is very far from being reached. In particular, there are no examples unambiguously demonstrating the role of transposon insertions in speciation, although there were many speculations on this point (see, for example, Hamdi et al., 2000).

The only way to reach a comprehensive understanding is by comparing various genome structures and revealing the correlation of phenotypic novelties and genetic changes, including acquisition of new transposable elements in new genomic sites. In principle, a very fast accumulation of

Abbreviations: HERV(s), human endogenous retrovirus(es); LINE(s), long intersparsed elements; LTR(s), long terminal repeat(s); Mya, million years ago; Myr, million years

* Corresponding author. Tel.: +7-095-330-6992; fax: +7-095-330-6538.
E-mail address: yuri@humgen.siobc.ras.ru (Y.B. Lebedev).

0378-1119/01/$ - see front matter
PII: S0378-1119(01)00570-4

sequencing information currently allows one to carry out detailed intergenomic comparisons by means of bioinformatic tools. However, unavoidable gaps and errors in the whole genomic sequences as well as a limited number of the available structures, in particular of the primate genomes, still make necessary direct experimental analysis of differences between genomes including positions of transposable elements.

We are interested in the possible impact of endogenous retroviruses (ERVs), which are one of the important representatives of transposable elements, on the primate evolution. The retrovirus-related structures are contained in the genomes of all the vertebrates investigated so far (Herniou et al., 1998). Some of the ERVs present in the human genome, for example ERV-L, have related sequences in genomes of various mammalians (Benit et al., 1999). Other HERVs are primate genomes specific with the integration times of <10–60 Myr (for references see Sverdlov, 2000). The germ-line integrated viral sequences were subsequently inherited in a stable Mendelian fashion and propagated due to retroposition events. In the human genome, ERV sequences are distributed over all chromosomes constituting up to 4.64% of the sequenced DNA (up to 8% if one includes MaLR elements in this list) (International Human Genome Sequencing Consortium, 2001). Some HERVs are present as provirus-like elements usually flanked by long terminal repeats (LTRs). They are severely mutated in the regions encoding the retroviral proteins. But much greater in number in the genome are solitary LTRs produced probably by recombinations between two LTRs of the same provirus and having no adjacent retroviral genes. Two characteristics of the LTRs make them important features of the genome: they contain regulatory elements and they probably tend to be integrated in transcriptionally active regions (Leib-Mosch and Seifarth, 1995). In particular, HERV-K (HML-2) LTR sequences contain putative hormone response elements, enhancers, promoters, polyadenylation signals and transcription factor binding sites which might be involved in the regulation of adjacent genes if inserted in an appropriate context (Seifarth et al., 1998; Sverdlov, 2000). On chromosome 19, it has been shown that HERV-K LTRs frequently lie in proximity to zinc finger and zinc finger-like genes. There are reports suggesting that LTRs modulate expression of nearby genes directly through transcription regulatory signals or by regulation of translation (for details and references see Brosius, 1999; Sverdlov, 2000; Kowalski et al., 1999). LTRs might also affect the gene expression indirectly, through DNA methylation and chromatin remodeling (for reference see Kass et al., 1997; Sverdlov, 2000). It should be noted that in the reported examples of the LTR involvement in gene regulation these elements are always incorporated in the 5′ proximal regions of genes. There is no doubt that many of the LTRs implementing regulatory functions (e.g. enhancers or silencers) in positions distant from the point of transcription initiation do exist but are still not identified. Therefore, to understand the role of the LTRs in the primate evolution one has to perform a whole genome comparison of the LTR positions in various primates followed by functional analysis of those LTRs that are differently located in the genomes under comparison. Inasmuch as there are tens of thousands of various LTRs and their parts in the human genome it is reasonable to restrict the research first by the LTRs that seem to be the best candidates for functional activity. We used three criteria to choose these LTRs: (i) the LTRs should be full-sized to comprise all regulatory elements; (ii) they should belong to a most biologically active HERV family, HERV-K; and (iii) they should be located outside of the clusters of interspersed repeats, because these clusters are most probably not functional. Despite all the limitations of such an approach, it can reveal candidate regulatory elements acquired in the course of evolution. We applied it to the human genome, studying chromosome by chromosome, first by mapping LTRs that meet the above-mentioned criteria and then by analyzing the presence of orthologous LTRs in the genomes of various primates. We have previously examined human chromosome 19 in a similar way. In this report, the object of our research was recently sequenced human chromosome 21. The chromosome sequences allowed us to make a refinement of our independent mapping of some LTRs. A phylogenetic analysis was performed for the LTRs located in the sites free of other repeats, and the time of their appearance in the primate genomes was determined. Moreover, human-specific LTRs were identified.

2. Materials and methods

2.1. Oligonucleotide primers

Oligonucleotide primers for PCR amplification and hybridization probes were synthesized using a Milligen 7500 DNA synthesizer. LTR-related primers were designed using multiple alignment of HERV-K LTR consensus sequences (Lavrentieva et al., 1998). Sequences of a set of suppression primers and an adapter used for specific PCR amplification of LTR-flanking sequences can be seen in Lebedev et al. (2000). Primers for genomic PCR of human and primate DNAs are shown in Table 1.

2.2. Cosmid and YAC libraries screening

A human chromosome 21 specific cosmid library LL21NC02 (Soeda et al., 1995) was spotted on high-density filters. LTR probes were prepared by PCR amplification of the total human DNA using a 19-for primer (GAGATCAGA(C/T)TGTTACTGTGTC) and an ltr-rev primer (AAAGACACAGAGACAAAGTATAGAG). Probe preparation and hybridization were as described previously (Lavrentieva et al., 1998). DNA from positive clones was prepared using a Wizard Plus Minipreps System (Promega). A minimal tiling path of YACs spanning 21q (Gardiner et al.,

Table 1
Primers for genomic PCR

Loci-specific primers	Sequence (5′–3′)
AP001037 R1	TGAACTATGTTTTCGGCTCTGA
AP001037 F1	TATTGCCAGTTCATCTCTCCAA
AP001037 F2	CAAACACACAGAAGCCATGT
AP000431 F1	GTTGGTTTGGTTTTACCCTC
AP000431 F2	CTCTTATCTGGATTATTGGA
AP000431 R1	TTGAATGTTGTAGATAAATAGGT
AL109763 F1	TCTTGCAAAGAATTCATGTTCAGT
AL109763 F2	TTGTTGCCTCCATGATACCC
AL109763 R1	TGTTCTGGAAACTATGGGCA
AL109763 R2	TCCACTGTGCCAGAACAACTG
AP000432 F1	ACCCGCTTGCGTTACCAATATC
AP000432 R1	ACGAGATAACCAGCCCACTTCC
AB005612 F1	CCAGTGCCACAAGGTCAG
AB005612 F2	CCGATTCCCCATTCATTCCAG
AB005612 R1	AAGAATGGCAGCGTTGATG
AB005612 R2	GTTGATGCCTGTCCCTCTGCC
AC006684 F1	TTGGGATGACCAGTAACCG
AC006684 F2	AGGGAACCAGCGCACACAGC
AC006684 R1	CATCTCTGGGCTAAGGCATC
AC006684 R2	TCAGTCCCACAAAGGCATCAGT
Q39E10 F1	GTGCGGAGGCGGTCTGCCTAGAAT
Q39E10 R1	TCTTTGTCCCTCTGCTTCCCAGTGGTT
Q21F12 F1	CAACACAGACTGCATAATGGTTAG
Q21F12 R1	GACATGTCCTCCATTTTCAGGCTAG
Q36G12 F1	GTTCATGCAACATCAGACCTCGT
Q36G12 R1	AACTTCACCCTAGAGAAAAGCCT
AP000041 F1	CAGGGCCAGGATTTGAAC
AP000041 R1	CCTGGCATACAACACTTAACG
AC006556 F1	GACTCCTCTTTCTCTTGCCATT
AC006556 R1	CGTGGTATCCCAAATTGAGC

1995) was also used in several cases for screening with an LTR probe.

2.3. Amplification of LTR sequences

LTR sequences were amplified by PCR using cosmid DNAs in 25 μl reaction volumes containing 0.2 μM each of the LTR-specific primers, 200 μM dATP, dTTP, dGTP, and dCTP, 2 mM $MgCl_2$, 0.5 units of AmpliTaq, and 2 ng of cosmid DNA. The amplifications were carried out in 25 cycles of 20 s at 95°C, 30 s at 60°C, and 45 s at 72°C.

2.4. Cosmid analysis

Cosmid DNA (150 ng) was digested with *Eco*RI, electrophoresed in 0.8% TAE agarose and transferred to Zeta-probe membranes (Bio-Rad). Filters were hybridized under standard conditions with an LTR probe obtained by PCR amplification either from the total human DNA or from a particular cosmid clone using the 19-for and ltr-rev primers and labeled with [α-^{32}P]dATP using a Prime-a-Gene Labeling System (Promega). Fluorescent *in situ* hybridization (FISH) of cosmid DNAs to human chromosomes was carried out by standard procedures.

2.5. Isolation of LTR-flanking regions

Selective suppression PCR was used for the preparation of LTR-flanking regions as described recently (Lebedev et al., 2000). Cosmids were digested with *Alu*I and ligated to suppression adapters. The PCR fragments obtained with LTR-specific and adapter primers were purified and sequenced manually with an *fmol* Sequencing System (Promega) using primers labeled with [γ-^{32}P]ATP by polynucleotide kinase.

2.6. Sequence analysis

Sequences were analyzed for homology using BlastN (available at http://www.ncbi.nlm.nih.gov/BLAST) and for repeat content using RepeatMasker2 (available at http://ftp.genome.washington.edu). Sequences were aligned using ClustalW (available at http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/clustalw.html) and PHYLIP (Phylogeny Inference Package) version 3.5c (Felsenstein, 1993). Construction of a neighbor-joining unrooted tree was carried out by aligning the LTR sequences using ClustalW followed by visualization using TreeView 1.6 (available at http://taxonomy.zoology.gla.ac.uk/rod/rod.html).

3. Results

3.1. Isolation and mapping of LTR sequences

Screening of the chromosome 21-specific cosmid library with the genome-derived LTR probe revealed 75 positive clones. To identify a subset containing complete LTR sequences, each cosmid was analyzed by PCR using primers corresponding to the 3′ and 5′ ends of the consensus LTR sequence. PCR amplification with 23 of 75 cosmids led to the formation of products with the expected length (~800 bp). Only these 23 clones were taken for further analysis. Overlaps among the 23 clones were determined by several methods: by searching for identical clone names in the maps previously constructed from the same library (Soeda et al., 1995; Yaspo et al., 1995), by *Eco*RI fingerprint analysis, and by FISH mapping. Four sets of overlapping clones and five unique cosmid clones were identified. Three of the contigs and four separate clones were subsequently mapped within 21q (LTRs 3, 5, 10, 12, 13, 14, and 15 in Fig. 1B). Sequences of the cosmid clones adjacent to the LTRs were determined and compared to those deposited in databases (see below).

A BlastN search of GenBank and a search of a recently created Chr21 database (Hattori et al., 2000) using the consensus HERV-K (HML-2) LTR sequences (Lavrentieva et al., 1998) identified matches to 12 segments of the chromosome 21 genomic DNA. The accession numbers and LTR positions on a 21q metric map (Hattori et al., 2000) are given in Table 2. For each of these 12 segments computer-calculated sizes of the *Eco*RI fragments surrounding the

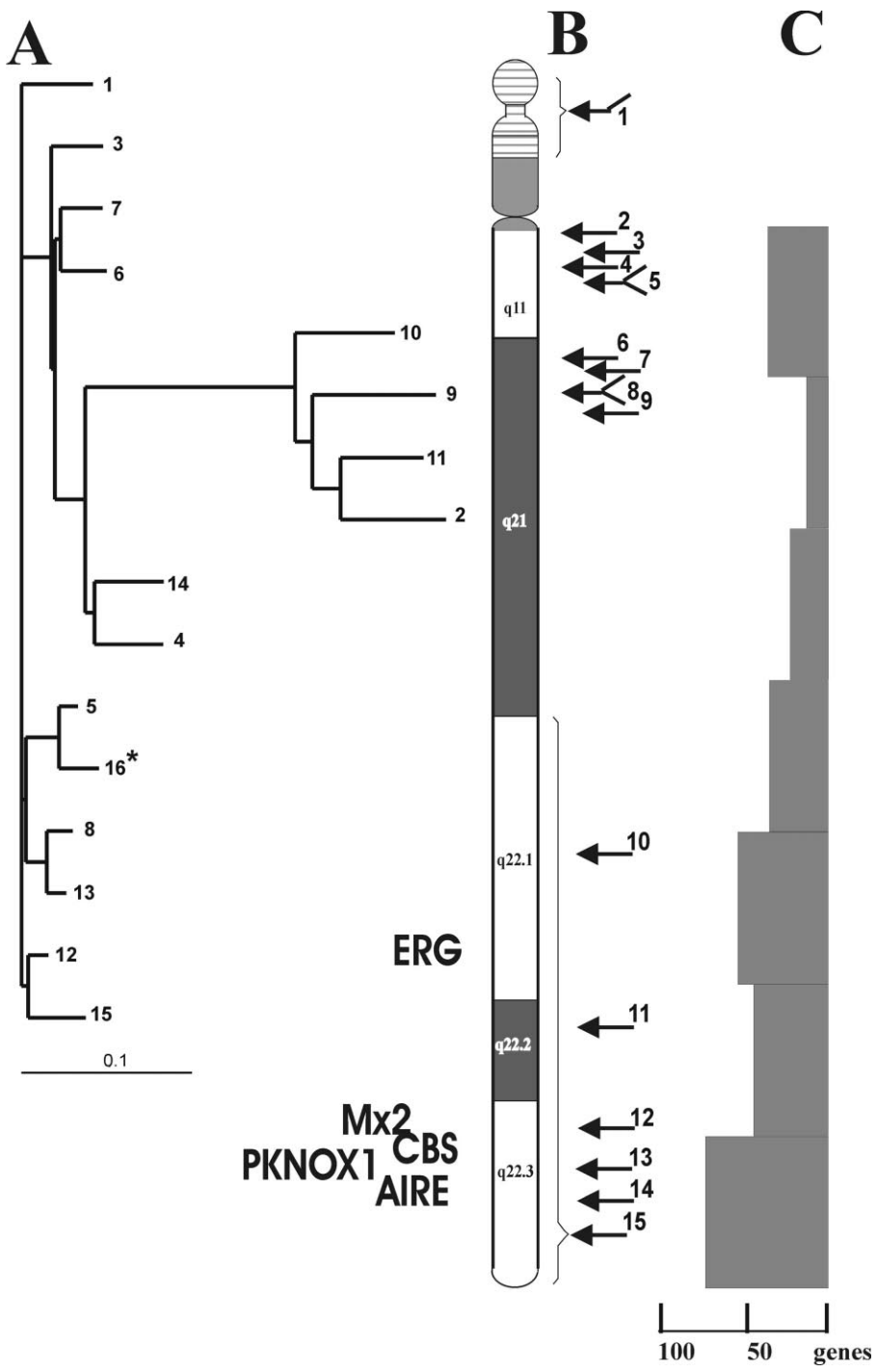

Fig. 1. A HERV-K LTR nearest neighbor dendrogram (A) and an ideogram of human chromosome 21 (B) with LTR locations and (C) genes density. Arrows mark positions of the mapped LTR sequences. Positions of HERV-K elements containing two LTRs are indicated with split arrows. Numbers at the arrows correspond to the following cosmids and GenBank Accession numbers: 1, Q36G12, AF260249; 2, AP001657; 3, Q21F12, Q4A12, Q89H11, AL109748; 4, AC006556; 5, Q94C9, Q7B6, Q91F3, Q42G10, Q62F6, Q34G2, AF260253; 6, AP000432; 7, AF165175; 8, AL109763; 9, AP000431; 10, Q76B8, AP000041; 11, AP001037; 12, Q14C10, Q58E2, 91E11, 39A4, 87A1, AC005612; 13, Q39E10, AP001631; 14, Q87A5, Q7C3, Q68E7, Q8G12, Q78D10, AB006684; 15, Q108B1, AF260251; 16*, Q37G12, AF260250. LTRs 1 and 15 were assigned to the chromosome bands marked by braces by FISH analysis of the corresponding cosmid clone. An LTR-containing cosmid Q37G12 was not assigned to any definite locus of Chr21; it is marked by an asterisk in the neighbor-joining tree. The diagram in (C) displays gene contents for seven regions each of 4.8 Mb in length. A total of 225 genes (categories 1–4) described for 21q (Hattori et al., 2000) were used for the calculation.

Table 2
Human genes in the vicinity of the LTRs

LTR		Neighboring genes		
#/Accession number	Position on Chr21 metric map (kb)[a]	Name (category)[a]	Orientation[b]	Distance (kb)[c]
1/AF260249	21p[d]	ND		
2/AP001657	57.9–58.9	*PRED65* (2.2)	−	71.6 (5′)
3/AL109748	858.4–859.4	*CNN2P* (5), *C21orf15* (4.2)	−, −	5.8 (5′), 22.2 (3′)
4/AC006556	1320.5–1321.5	*PRED6* (3.1), *RBM11* (3.1)	−, −	Intron 2, 53.5 (3′)
5/AF260253	21q11.1[d]	ND		
6/AP000432	4743.9–4744.9	*BTG3* (1.1), *YG81* (1.1)	+, +	81.4 (5′), 99.1 (3′)
7/AF165175	4983.4–4984.4	*RL37P* (5), *C21orf39* (4.2)	−, +	39.2 (3′), 48.1 (5′)
8[e]/AL109763	5614.1–5622.5[e]	None		
9/AP000431	6158.6–6159.6	*SLC6A6P* (5)	−	> 140 (5′)
10/AP000041	20003.2–20004.2	*PRKCBP2* (1)	+	30.1 (5′)
11/AP001037	25596.2–25597.2	*ERG* (1.1)	+	Intron 3
12/AC005612	28449.1–28450.1	*MX1* (1.1), *MX2* (1.1)	−, −	4.4 (5′), 12.0 (3′)
13/AP001631	30114.6–30115.6	*PKNOX* (1.1), *CRYAA* (1.1)	−, −	9.0 (3′), 20.8 (5′)
14/AB006684	31210.5–31211.4	*AIRE* (1.1), *DNMT3L* (1.2)	+, −	2.4 (5′), 10.0 (5′)
15/AF260251	21q22[d]	ND		
16/AF260250	ND	ND		

[a] Positions of the LTRs and categories of the genes are presented in accordance with the Chr21 database (Hattori et al., 2000).
[b] ‘+’ and ‘−’ designate same and opposite directions of transcription, respectively.
[c] 5′ or 3′ in parentheses indicate the end of the gene which is nearest to the LTR.
[d] Location of the corresponding cosmid clone determined by FISH.
[e] Borders of a defective HERV-K (HML-2) provirus are marked.

LTR sequences were compared with the experimental *Eco*RI restriction patterns for the cosmid clones described above. The accordance of the patterns was considered as evidence of the identity of the clone and the database sequences. As a result, five of our cosmid clones/contigs as well as LTR flanks in these cosmids sequenced by us were found to be identical to the corresponding GenBank/Chr21 sequences. The other four experimental clones/contigs were not found among chromosome 21 sequences. On the other hand, we could not detect seven LTRs and their flanks among the LTR-containing cosmids. The reason why these seven LTRs were not detected in our cosmid library is not clear; possibly the library was not complete. On the other hand, the detection of the LTR-containing cosmids absent from the sequences deposited (Hattori et al., 2000) deserves special attention. One of these LTRs (16 in Fig. 1A) was found in a cosmid clone not overlapping with others. Its position on the chromosome was not defined and at this point it can be assigned to the chromosome only tentatively. Two other LTRs (1 and 15 in Fig. 1B) were mapped to the chromosome by FISH within the non-sequenced 21p arm (clone 1) and 21q22 (clone 15) region containing three clone gaps. The last of the LTRs (5 in Fig. 1B) was detected in a contig of seven overlapping cosmids. As judged from the databases available, none of the clones were sequenced. Also a YAC 849B10 (Gardiner et al., 1995) located near the ACEM breakpoint (Hattori et al., 2000) in the 21q11.1 region was found to cross-hybridize with the cosmids from the contig. Therefore, we assigned this contig to the pericentromeric region of this chromosome. A detailed analysis allowed us to detect another LTR within this contig. We sequenced the LTRs with the adjacent regions but could not find perfect matches of these sequences to the available chromosome 21 sequences from this area. These LTRs and their surroundings probably also fall in a sequence gap. One of the LTRs was attached to a retroviral *env*-related sequence. The two LTRs (5 in Fig. 1B) were highly homologous (99.4% identity) but still different. According to the Genome Database, their genomic flanking sequences are perfectly identical to the sequences flanking a human specific full-sized HERV-K (HML-2) provirus (Barbulescu et al., 1999; Accession numbers: AF164616, AF164919, AF165235) not assigned to any human chromosome. The sequences of the LTRs were also nearly identical to those published by Barbulescu et al. (1999) (99.5% for 5′ LTR and 99.7% for 3′ LTR). Finally, Barbulescu et al. (1999) determined that this HERV is human, in accordance with our data. Therefore, we came to the conclusion that these two LTRs belong to the provirus described by Barbulescu et al. (1999) and that the provirus is located in the pericentromeric region of chromosome 21 as shown in Fig. 1B. Another proviral sequence was found in the Chr21 database within 21q21 (AL109763, 8 in Fig. 1B). It contains a full-size LTR at the 5′ end and a sequence corresponding to a part of the U3 region of the 3′ LTR. The LTR situated on the short arm (1 in Fig. 1) was attached to the primer binding site (PBS) and a *gag*-related sequence of the HERV-K (HML-2) proviruses, but no other LTRs were found in this cosmid clone. All of the other 12 LTRs mapped on the q arm (2–4, 6, 7 and 9–15 in Fig. 1B)

were characterized as solitary by the identification of short direct repeats of the genomic DNA at the LTR borders and by the absence of adjacent retroviral genes. Fifteen regions containing full-sized LTRs of the HERV-K (HML-2) family were thus mapped on chromosome 21.

3.2. Distribution of the LTRs along Chr21

The positions of the 15 mapped HERV-K (HML-2) elements are shown on the ideogram of human chromosome 21 presented in Fig. 1B. One of the 15 mapped elements (1 in Fig. 1B) is located at the acrocentric stalk of the p arm as shown by FISH using a corresponding cosmid clone (Q36G12) as a probe. Two other provirus-like structures were mapped on the q arm. Six of the LTRs are located within 21q22, and three of them were mapped to distal 21q22.3. Positions of the LTRs mapped on 21q relative to the locations of human genes are summarized in Table 2. As follows from Fig. 1, their distribution is not only non-uniform but also roughly correlates with previously described gene densities (Gardiner, 1997; Antonarakis, 1998; Hattori et al., 2000). It should also be noted that the pericentromeric region of about 1 Mb in length on the chromosome is composed of interchromosomal repeats (International Human Genome Sequencing Consortium, 2001). The occurrence of some LTRs (LTRs 1–5, Table 2) in this loci might be associated with rearrangements making the apparent preference of LTRs to GC- and gene-rich regions even more pronounced.

3.3. LTR sequence diversity

We have previously shown that currently known HERV-K LTR sequences can be divided into 16 distinct LTR groups (Lavrentieva et al., 1998; Lebedev et al., 2000) based on specific patterns of mutations and deletions. We have also demonstrated that some of these groups appeared as recently as 3–6 million years ago, whereas the others can be as old as 50 million years. A similar analysis carried out for 16 LTR sequences found on chromosome 21 shows that each of them falls into one of eight of the previously described groups, and that the evolutionary ages of the group master genes varied from 10 to 40 Myr. The results are summarized in Table 3.

3.4. Non-random LTR distribution along chromosome versus random alternation of various LTRs

A pairwise comparison of the 16 LTR sequences (only one of two #5 provirus LTRs was taken for the comparison) showed that the number of differences ranged from as low as 14 nucleotides (around 1.5% of an average LTR length of 971 bp) to 269 nucleotides (over 30%). The neighbor-join-

Table 3
Phylogenetic assessment of the integration times for individual HERV-K LTRs in the primate genomes

LTR on Chr21/groups[a]	Primate species[b]							Integration time, Myr
	Hu	Ch	Gor	Oran	Gib	OWm	NWm	
13/II-La	+	−	−	−	−			< 5.6
8[c]/II-Lb	+	− [d]	?	?	?	?	?	< 5.6
1/II-T	+	+	?	?	?	?	?	5.6–8
5[c,e]/II-T	+	+	?	?	?	?	?	5.6–8
3/II-O	+	+	+	?	−	−	?	8–13
12/II-T	+	+	+	−	−	?	?	8–13
6/II-O	+	+	+	−	?	?	?	8–13
14/II-O	+	+	+	+	−	−	?	13–18
9/I-S	+	+	+	+	+	−	−	18–28
4/I-X	+	+	+	+	+	−	?	18–28
11/I-E	+	+	+	+	+	+	?	> 28
10/I-Ka	+	+	+	+	+	+	?	> 28
15[f]/II-T	ND	ND	ND	ND	ND	ND	ND	~ 15[g]
16[f]/II-T	ND	ND	ND	ND	ND	ND	ND	~ 15[g]
7[f]/II-O	ND	ND	ND	ND	ND	ND	ND	~ 16[g]
2[f]/I-S	ND	ND	ND	ND	ND	ND	ND	~ 27[g]

[a] The LTR numbering corresponds to that in Fig. 2.

[b] +, successful PCR amplification; −, presence of a short PCR product with the length corresponding to the site lacking the LTR; ?, no PCR fragment detected. The branching data for primate evolution were averaged from three estimates (for references see Lebedev et al., 2000): New World monkeys, 45 Myr; Old World monkeys, 28 Myr; Gibbon, 18 Myr; Orangutan, 13 Myr; Gorilla, 8 Myr; and Chimpanzee, 5.6 Myr.

[c] Presence of the HERV-K element detected by PCR with a primer corresponding to a unique genome sequence and the second primer corresponding to a part of the *gag* retroviral gene. The lack of the PCR product in these cases was interpreted as the absence of the corresponding HERV-K.

[d] A short PCR fragment was detected using amplification with a pair of primers corresponding to the human genome region surrounding the HERV-K provirus integration site.

[e] The results correspond to Barbulescu et al. (1999).

[f] The LTR sequences in human loci are attached to repeats which prevents unique primer design.

[g] Predicted time of the insertion calculated using the degree of divergence from the corresponding group consensus.

ing method was used to calculate the degree of similarity of the LTRs (Fig. 1A). A comparison of the neighbor-joining dendrogram with the LTR map of the chromosome (Fig. 1B) revealed no correlation between the physical neighborhood of the LTRs and the degree of their similarity. Pairs of the LTRs with the highest levels of identity on the dendrogram were often located at large distances from each other on the chromosome. For example, sequences 8 and 13 are highly similar but are located close to the centromere and in 21q22.3, respectively. Distant reciprocal location is characteristic of sequences 4 and 14 and other pairs of sequences.

3.5. Individual LTR evolutionary ages and maintenance in primates

To determine relative evolutionary ages of the LTRs, their availability in a number of primate species was tested. To do this, primers were designed to unique sequences flanking each LTR in the human DNA and used for PCR amplification from orthologous loci in the genomes of human, chimp, gorilla, orangutan, gibbon, and various species of Old World and New World monkeys. In each case, the identity of the PCR product was verified by hybridization both with LTR-specific probes and primate-specific probes free of the LTR sequences. As shown in Fig. 2 and summarized in Table 3, this revealed different patterns of presence and absence of particular LTRs in the genomes studied. Failure of some PCR amplifications could be due to the primers' target sequence divergences, and therefore no certain conclusions on the presence or absence of the LTRs in a given species could be made. However, a number of positive amplifications allowed us to determine the age of several insertions based on previously reported branching data for the primate evolution. In most cases, individual LTRs were found to be younger than the corresponding master genes (see Lebedev et al., 2000) suggesting prolonged activity of the master genes.

4. Discussion

In this research, we determined the precise location on human chromosome 21 of a subset of HERV-K LTRs selected as the most probable candidates for being functional. Although they represent only a small fraction of all LTR elements found on the chromosome, we believe that the features of this subset can be typical for the retroviral remnants domesticated by the genome for its own functional purposes (see below).

4.1. The distribution of the HERV-K (HML-2) LTRs along chromosome 21 is uneven and roughly correlates with the gene distribution

It is becoming increasingly clear that the gene regulation in mammalian genomes involves enormously complex networks of *cis*-regulatory elements interacting with *trans*-acting factors. The regulatory systems include different layers of regulatory information necessary to achieve and maintain correct tissue and developmental specificity of the gene expression. The *cis*-regulatory sequences can be physically very distant from the genes under their control. Sometimes the sequences responsible for correct spatial and temporal regulation of a particular gene locus are scattered over hundreds of kilobases of DNA (for brief review see Bonifer, 2000). Since LTRs are obvious candidate sequences for being a part of the genome regulatory machinery, we attempted to correlate the positions of the LTRs and genes within the chromosome. Fig. 1C shows that the density of the LTR sequences subset on chromosome 21 roughly correlates with the known, non-uniform density of genes (see Fig. 1C and Hattori et al., 2000). Two clusters of the LTRs were observed, one of them located in the centromere proximal region, and another one in the telomere proximal region. Such a clustered distribution is in striking contrast with the almost even distribution of the same type of LTRs on human chromosome 19 (Lavrentieva et al., 1998) where genes are also distributed evenly. Thus, the LTRs in both cases tend to neighbor genes along the chromosome. The distribution of the LTRs is consistent with their predominant integration in transcriptionally active regions of the genome (Leib-Mosch and Seifarth, 1995; Sverdlov, 2000). It probably reflects the trend of retroviruses to be integrated into open chromatin regions (for review see Rynditch et al., 1998). It is also interesting to correlate the distribution of LTRs with the GC content along the chromosome DNA. It has been earlier reported that HERVs integrate predominantly into "GC- and Alu-rich, actively transcribed and early replicating chromatin regions" (Leib-Mosch and Seifarth, 1995). After the human genome sequence publication it became clear that LTRs are rather uniformly distributed among sequences of various GC content being relatively less abundant only in the most GC-rich regions, whereas LINEs occur at much higher density in AT-rich loci, and Alu sequences prefer GC-rich DNA (International Human Genome Sequencing Consortium, 2001). On chromosome 21 the telomere proximal cluster of the HERV LTRs resides within chromosome DNA of as high as about 50% GC content, whereas the centromere proximal cluster is located within a relatively GC poor region (~40%), although local increases of the GC content in the sites of the LTR can occur. Therefore, it might be cautiously concluded that it is the gene density rather than the GC content that determines where the LTRs will be predominantly integrated. However, this very preliminary conclusion remains to be confirmed by analysis of more data accumulated in the course of the human genome sequencing.

4.2. LTR–gene relations

At present it is very difficult (if at all possible) to establish functional relations between a certain gene and a given LTR

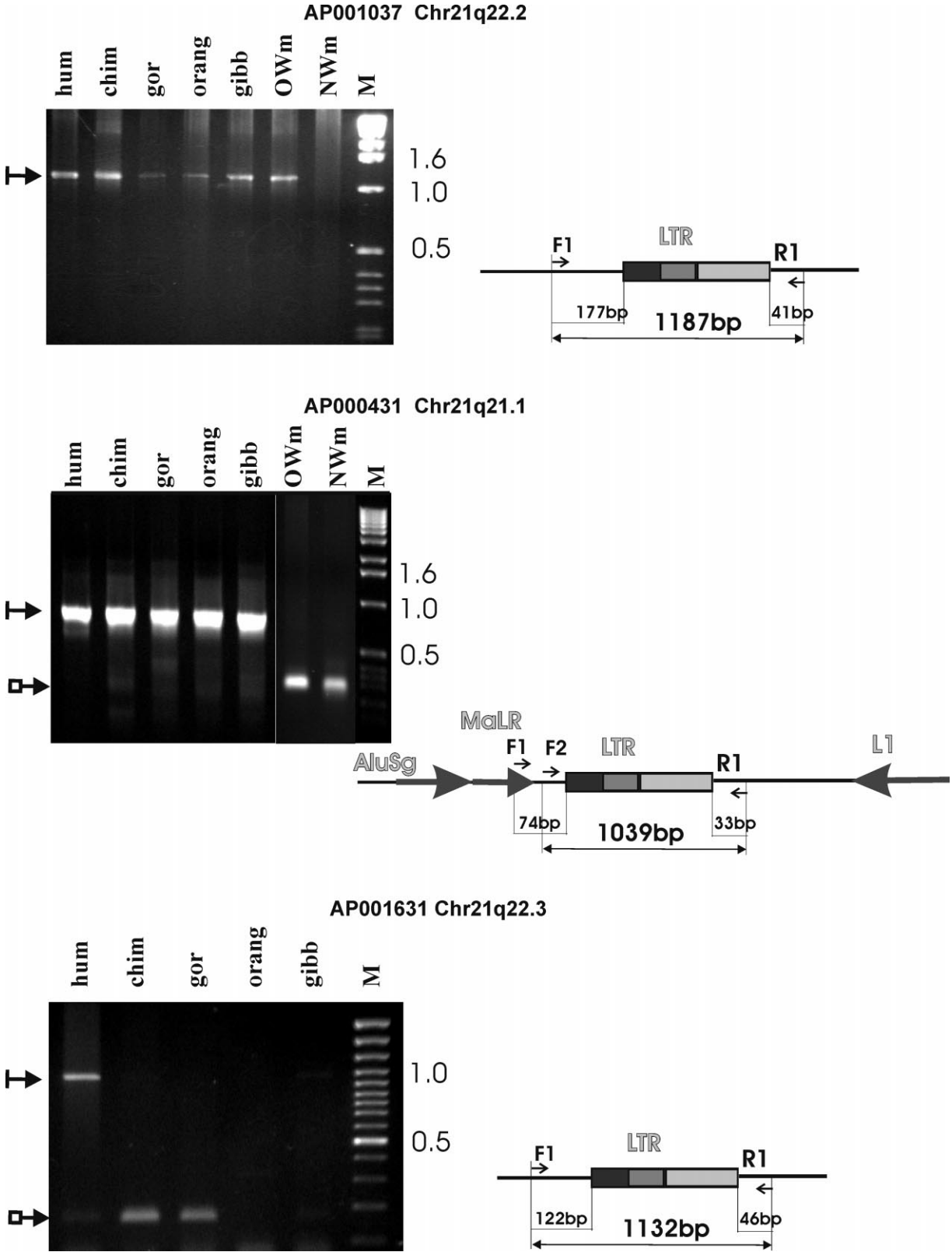

Fig. 2. The results of three individual LTR-containing loci PCR amplifications in the human and other primate genomes. The schemes of the loci are given together with corresponding electrophoregrams of the PCR products in agarose gels. Arrows to the left of the electrophoregrams indicate predicted locations of the LTR-containing and LTR-less PCR products. The schemes of human loci show positions of the LTRs (gray boxes), other repeated elements (bold arrows), and specific primers (F, forward; R, reverse) used for the PCR amplifications. Expected lengths of PCR products are marked below the loci schemes.

on chromosome 21 because of its 225 'genes' only 127 are 'true' genes, while the rest are rather putative genes. Therefore, the correlation can only be approximate. However, we considered it to be useful to try to order the LTRs relative to all kinds of genes. The results of such an ordering are presented in Table 2. It can be seen that most of the LTRs mapped on the long arm of chromosome 21 are located so that at least one of the genes lies within reach of their potential enhancer action. For example, LTR 14 (Fig. 1B) is located 2.38 kb upstream of the first exon of the *AIRE* (autoimmune regulator, APECED protein) gene, and two of the LTRs are situated within introns of the genes (Table 2). These LTRs might serve as transcriptional enhancers for the corresponding genes. The LTRs more distant from the genes may still be involved in the regulation through large distance mechanisms which now attract increasing attention. These mechanisms include locus control regions (LCRs) (Long et al., 1998), chromatin remodeling, and synthesis of non-coding RNAs further acting as regulators like, for example, RNAi (for a review see Eddy, 1999). Certainly, the correlations observed by no means imply functional relations but they can give one of the directions for future functional assays. Some of the LTRs in Table 2 are known to be transcriptionally active (Vinogradova et al., unpublished data).

Most of the LTRs reported so far as components of gene regulatory systems were positioned in 5′ upstream regulatory areas of the genes (for review see Brosius, 1999). Clearly, the more distant the LTR is from the gene the more difficult it is to identify the LTR as an element essential for the gene expression. An example of such a long distance LTR possibly involved in the regulation of human beta-like globin genes in erythroid cells was reported recently (Long et al., 1998). The authors found an LTR retrotransposon belonging to the ERV-9 family of HERVs in the apparent 5′ boundary area of the LCR at about 5 kb from the genes. The LTR possesses enhancer activity and may possibly serve a relevant function in regulating the transcription of the beta-globin LCR. Many other LTRs retaining their regulatory potential might similarly be involved in regulation. Therefore, the study of individual LTRs and, in particular, those situated wide apart from the genes (we will call them 'orphan LTRs') can be expected to reveal many new regulatory elements in different parts of

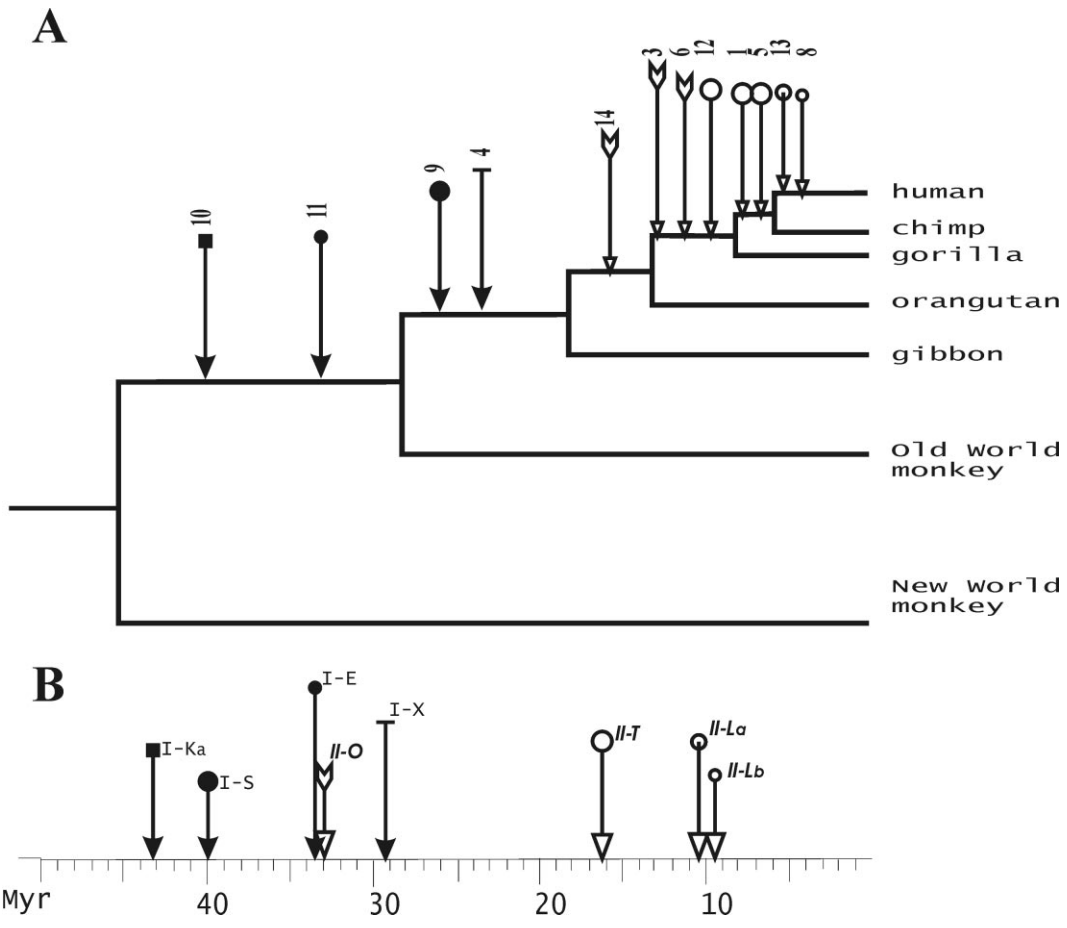

Fig. 3. Integration times of individual HERV-K elements mapped on Chr21. (A) (Upper) An evolutionary tree of the primate lineage. Arrows mark the times of the individual LTR insertions. Accession numbers or cosmid clone names of the LTR-containing sequences are added to the arrows. Top ends of the arrows are specific for each of the LTR groups. (B) Insertion times of the master genes having formed the groups of the LTRs mapped on chromosome 21. Alphanumeric marks over the arrows denote names of the LTR groups (Table 2).

the genome. The data like those compiled in Table 2 can help in the identification of such long distance regulators.

4.3. The intra- and interchromosomal distribution of the LTRs

No correlation between the physical positions of LTRs on chromosome 21 and the level of their identity or their integration times was observed (Fig. 1 and Table 3). The same random type of LTRs distribution was previously reported by us for human chromosomes 19 and 22 (Lavrentieva et al., 1998; Artamonova et al., 2000). It should be noted, however, that the LTR distribution among chromosomes is non-random: there are both human chromosomes enriched in LTRs and 'LTR-poor' chromosomes (Leib-Mosch and Seifarth, 1995). The explanation of this apparent contradiction will also help to elucidate mechanisms of integration of LTRs, their evolutionary maintenance, and the roles they might play in the regulation of gene expression.

4.4. LTRs of different ages are present on chromosome 21, but relatively young LTRs are more abundant

The analysis presented here demonstrates the evolutionary history of the appearance of the HERV-K LTRs in the human and primate genomes, illustrated in Fig. 3. Two LTRs (8 and 13) appeared in the human genome after the divergence of the human and chimpanzee lineages and are human-specific, two others (1 and 5) appeared after the divergence of the gorilla lineage from the human–chimp common ancestor, and the next three (3, 6 and 12) appeared after the orangutans split. Only five (31%) of the 16 LTRs have integration times above 18 Myr. This proportion is considerably smaller than the percentage (72%) of the LTRs older than 18 Myr among 383 HERV-K (HML-2) LTRs found by us in the non-redundant NCBI Genome Database. Again, what does this selectivity stem from? Are there any parallels with non-random distribution of the LTRs among chromosomes? If we accept a standard concept of retropositions that is "transcription – transport of the transcript in cytoplasm – reverse transcription – transport of cDNA in nucleus and reintegration" (Lower et al., 1996), then why should the newly synthesized cDNA prefer some chromosomes or even particular regions of chromosomes for the integration? Or maybe randomly integrated LTRs are then specifically deleted from some chromosomes or chromosome areas? These questions remain to be answered, and to do this additional chromosomes should be analyzed to accumulate more data on the specificity of the distribution of the LTRs.

Acknowledgements

The authors thank Dr B.O. Glotov for fruitful discussions and help in the manuscript preparation and Dr V.K. Potapov for oligonucleotide synthesis. The work was supported by grants of the Human Genome State Project of Russia, RFBR 98-04-48798, INTAS-99-01143 and HHMI International Research Scholar's award 75195-544201.

References

Antonarakis, S.E., 1998. 10 years of genomics, chromosome 21, and Down syndrome. Genomics 51, 1–16.

Artamonova, I.I., Gorodentseva, T.N., Lebedev, Y.B., Sverdlov, E.D., 2000. Nonrandom distribution of the endogenous retroviral regulatory elements HERV-K LTR on human chromosome 22. Dokl. Biochem. 372, 87–89.

Barbulescu, M., Turner, G., Seaman, M.I., Deinard, A.S., Kidd, K.K., Lenz, J., 1999. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr. Biol. 9, 861–868.

Benit, L., Lallemand, J.B., Casella, J.F., Philippe, H., Heidmann, T., 1999. ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. J. Virol. 73, 3301–3308.

Bonifer, C., 2000. Developmental regulation of eukaryotic gene loci: which cis-regulatory information is required? Trends Genet. 16, 310–315.

Britten, R.J., 1997. Mobile elements inserted in the distant past have taken on important functions. Gene 205, 177–182.

Brosius, J., 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238, 115–134.

Dunham, I., et al., 1999. The DNA sequence of human chromosome 22. Nature 402, 489–495.

Eddy, S.R., 1999. Noncoding RNA genes. Curr. Opin. Genet. Dev. 9, 695–699.

Felsenstein, J., 1993. PHYLIP version 3.5c. Distributed by the author. Based program in Felsenstein, J., 1989. PHYLIP – Phylogeny Inference Package. Cladistics 5, 164–166.

Gardiner, K., 1997. Clonability and gene distribution on human chromosome 21: reflections of junk DNA content? Gene 205, 39–45.

Gardiner, K., Graw, S., Ichikawa, H., Ohki, M., Joetham, A., Gervy, P., Chumakov, I., Patterson, D., 1995. YAC analysis and minimal tiling path construction for chromosome 21q. Somat. Cell Mol. Genet. 21, 399–414.

Hamdi, H.K., Nishio, H., Tavis, J., Zielinski, R., Dugaiczyk, A., 2000. Alu-mediated phylogenetic novelties in gene regulation and development. J. Mol. Biol. 299, 931–939.

Hattori, M., et al., 2000. The DNA sequence of human chromosome 21. Nature 405, 311–319.

Herniou, E., Martin, J., Miller, K., Cook, J., Wilkinson, M., Tristem, M., 1998. Retroviral diversity and distribution in vertebrates. J. Virol. 72, 5955–5966.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Kass, S.U., Pruss, D., Wolffe, A.P., 1997. How does DNA methylation repress transcription? Trends Genet. 13, 444–449.

Kazazian, H.H., 2000. Genetics. L1 retrotransposons shape the mammalian genome. Science 289, 1152–1153.

Kowalski, P.E., Freeman, J.D., Mager, D.L., 1999. Intergenic splicing between a HERV-H endogenous retrovirus and two adjacent human genes. Genomics 57, 371–379.

Lavrentieva, I., Khil, P., Vinogradova, T., Akhmedov, A., Lapuk, A., Shakhova, O., Lebedev, Y., Monastyrskaya, G., Sverdlov, E.D., 1998. Subfamilies and nearest-neighbour dendrogram for the LTRs of human endogenous retroviruses HERV-K mapped on human chromosome 19: physical neighbourhood does not correlate with identity level. Hum. Genet. 102, 107–116.

Lebedev, Y., Belonovich, O., Zybrova, N., Khil, P., Kurdyukov, S., Vinogradova, T., Hunsmann, G., Sverdlov, E., 2000. Differences in HERV-K LTR insertions in orthologous loci of human and great apes. Gene 247, 265–277.

Leib-Mosch, C., Seifarth, W., 1995. Evolution and biological significance of human retroelements. Virus Genes 11, 133–145.

Long, Q., Bengra, C., Li, C., Kutlar, F., Tuan, D., 1998. A long terminal repeat of the human endogenous retrovirus ERV-9 is located in the 5′ boundary area of the human beta-globin locus control region. Genomics 54, 542–555.

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.

Rynditch, A., Zoubak, S., Tsyba, L., Tryapitsina-Guley, N., Bernardi, G., 1998. The regional integration of retroviral sequences into the mosaic genomes of mammals. Gene 222, 1–16.

Seifarth, W., Baust, C., Murr, A., Skladny, H., Krieg-Schneider, F., Blusch, J., Werner, T., Hehlmann, R., Leib-Mosch, C., 1998. Proviral structure, chromosomal location, and expression of HERV-K-T47D, a novel human endogenous retrovirus derived from T47D particles. J. Virol. 72, 8384–8391.

Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. 9, 657–663.

Soeda, E., Hou, D.-X., Osoegawa, K., Atsuchi, Y., Yamagata, T., Shimokawa, T., Kishida, H., Soeda, E., Okano, S., Chumakov, I., Cohen, D., Raff, M., Gardiner, K., Graw, S.L., Patterson, D., de Jong, P., Ashworth, L.K., Slezak, T., Carrano, A.V., 1995. Cosmid assembly and anchoring to human chromosome 21. Genomics 25, 73–84.

Sverdlov, E.D., 2000. Retroviruses and primate evolution. BioEssays 22, 161–171.

Yaspo, M.-L., Gellen, L., Mott, R., Korn, B., Nizetic, D., Poustka, A., Lehrach, H., 1995. Model for a transcript map of human chromosome 21. Isolation of new coding sequences from exon and enriched cDNA libraries. Hum. Mol. Genet. 4, 1291–1304.