Insertional polymorphisms of full-length endogenous retroviruses in humans

Geoffrey Turner*, Madalina Barbulescu*, Mei Su*, Michael I. Jensen-Seaman[†‡§], Kenneth K. Kidd[†] and Jack Lenz*

Human endogenous retrovirus K (HERV-K) is distinctive among the retroviruses in the human genome in that many HERV-K proviruses were inserted into the human germline after the human and chimpanzee lineages evolutionarily diverged [1, 2]. However, all full-length endogenous retroviruses described to date in humans are sufficiently old that all humans examined were homozygous for their presence [1]. Moreover, none are intact; all have lethal mutations [1, 3, 4]. Here, we describe the first endogenous retroviruses in humans for which both the full-length provirus and the preintegration site alleles are shown to be present in the human population today. One provirus, called HERV-K113, was present in about 30% of tested individuals, while a second, called HERV-K115, was found in about 15%. HERV-K113 has full-length open reading frames (ORFs) for all viral proteins and lacks any nonsynomymous substitutions in amino acid motifs that are well conserved among retroviruses. This is the first such endogenous retrovirus identified in humans. These findings indicate that HERV-K remained capable of reinfecting humans through very recent evolutionary times and that HERV-K113 is an excellent candidate for an endogenous retrovirus that is capable of reinfecting humans today.

Addresses: *Department of Molecular Genetics, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA. †Department of Genetics and ‡Department of Anthropology, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06520, USA.

Present address: §Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA.

Correspondence: Jack Lenz
E-mail: lenz@aecom.yu.edu

Received: **31 July 2001**
Revised: **20 August 2001**
Accepted: **20 August 2001**

Published: **2 October 2001**

Current Biology 2001, 11:1531–1535

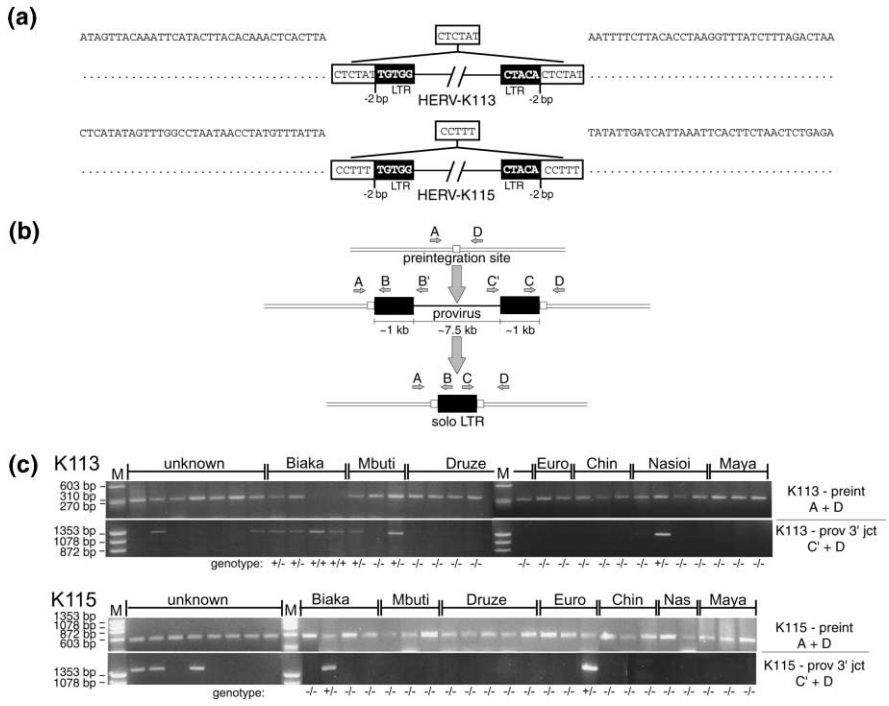
0960-9822/01/$ – see front matter

Results and discussion

BACs containing full-length HERV-K provirus clones were isolated as previously described [1] by screening filters from the human RP11 BAC library (BACPAC Resources) with a hybridization probe from the HERV-K *pol* gene. Since each provirus is inserted at a different position in the human genome, the DNA sequences flanking each provirus are unique to that particular provirus. The sequences immediately flanking both sides of each cloned provirus were determined. Two proviruses identified in this manner were called HERV-K113 and HERV-K115. A BLAST search of the nr and htgs databases in GenBank with the sequences flanking these proviruses identified only entries corresponding precisely to the empty or preintegration site alleles that lack HERV-K sequences (Figure 1a). The sequences were localized on the sequence assembly of the Human Genome Project Working Draft at UCSC for HERV-K113 at chromosome 19p13.11 and for HERV-K115 at chromosome 8p23.1. Proviral insertions were associated with duplication of a target sequence of 6 bp (CTCTAT) for HERV-K113 and 5 bp (CCTTT) for HERV-K115. Both 5 and 6 bp duplications have been observed previously for HERV-K [1]. In addition, 2 bp were deleted from the ends of the viral LTRs, unambiguously indicating that the proviruses were generated by a standard retroviral-integration process [5]. The existence of the proviruses and sequences corresponding to the preintegration sites suggested that both alleles exist at both loci in humans today.

To test this, a PCR strategy (Figure 1b) was used to search for both alleles in samples of human genomic DNA. Reactions using primers in the sequences flanking each provirus can distinguish the ancestral preintegration site, the full-length provirus, and the solo long terminal repeat (LTR) that can form by homologous recombination between the two LTRs of a provirus [1]. Genotyping of a small number of genetically diverse humans showed that HERV-K113 and HERV-K115 are each present in some humans but not others (Figure 1c). The human genomic DNAs tested included individuals of sub-Saharan African (Biaka and Mbuti), Middle Eastern (Druze), European, Chinese, Melanesian (Nasioi), and Native American (Mayan) origin, plus several placental DNAs of unknown ethnic origin. HERV-K113 was present in 9 of 31 individual samples (29%), including two Biaka individuals who were homozygous for the provirus allele. HERV-K115 was detected in 5 of 31 samples (16%). No solo LTRs were detected at either of these loci among the samples

Figure 1



(a) ATAGTTACAAATTCATACTTACACAAACTCACTTA CTCTAT AATTTTCTTACACCTAAGGTTTATCTTTAGACTAA
CTCTAT TGTGG CTACA CTCTAT
-2 bp LTR HERV-K113 LTR -2 bp
CTCATATAGTTTGGCCTAATAACCTATGTTTATTA CCTTT TATATTGATCATTAAATTCACTTCTAACTCTGAGA
CCTTT TGTGG CTACA CCTTT
-2 bp LTR HERV-K115 LTR -2 bp

(b) A D preintegration site
A B B' C' C D provirus
~1 kb ~7.5 kb ~1 kb
A B C D solo LTR

(c) K113 M unknown Biaka Mbuti Druze M Euro Chin Nasioi Maya
603 bp 310 bp 270 bp 1353 bp 1078 bp 872 bp
K113 - preint A + D
K113 - prov 3' jct C' + D
genotype: +/- +/- +/+ +/+ +/- -/- +/- -/- -/- -/- -/- -/- -/- -/- -/- -/- -/- -/- -/- -/- -/- -/-

K115 M unknown M Biaka Mbuti Druze Euro Chin Nas Maya
1353 bp 1078 bp 872 bp 603 bp 1353 bp 1078 bp
K115 - preint A + D
K115 - prov 3' jct C' + D
genotype: -/- +/- -/- -/- -/- -/- -/- -/- -/- -/- -/- -/- -/- +/- -/- -/- -/- -/- -/- -/- -/- -/-

Detection of HERV-K113 and HERV-K115 in human genomic DNA. **(a)** For each provirus, the top line shows the sequence of the preintegration site allele while the lower line shows the sequence of the proviral allele. The open box contains the target site that was duplicated upon proviral integration. The black box indicates the viral long-terminal repeat (LTR), with five nucleotides at the ends of the LTRs shown. Positions of the 2 bp deletions that occurred at the ends of the LTRs upon integration are indicated. **(b)** The PCR strategy used to detect the preintegration site, the provirus, and the solo LTR is shown. Black boxes represent the viral LTRs, and open boxes represent the target site duplications. Positions and orientations of the PCR primers (A, B, B′, C′, C, and D) are shown. The product generated with the AD primer pair is larger from the solo LTR than from the preintegration site by the size of a viral LTR (~970 bp). **(c)** PCR products obtained with the indicated primer pairs are shown. Only the proviral 3′ junction reactions are shown. The same results were obtained with the proviral 5′ junctions. The human population of origin for each sample is indicated: Euro, European; Chin, Chinese; Nas, Nasioi; M, size marker. The smaller band at the bottom of the HERV-K115 A + D reactions is unincorporated primers.

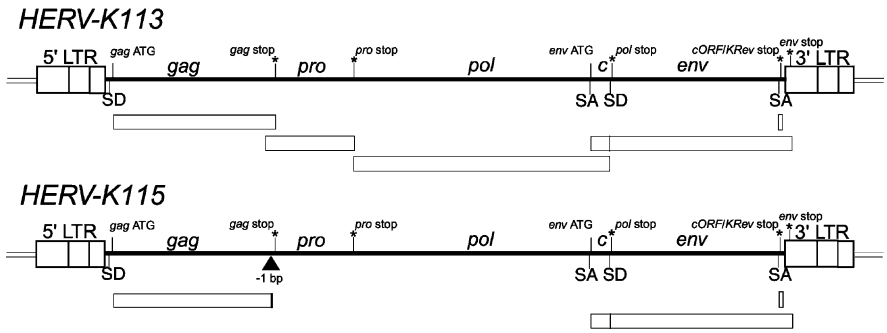
tested. Excluding the placental samples that may have contained three alleles, the HERV-K113 provirus allele frequency was 0.19 (9/48) in 24 diploid samples tested. There may be considerable variation among different human populations. The HERV-K115 provirus allele frequency was 0.04 (2/46) in 23 diploid samples tested. It is curious that the single anonymous donor for the RP11 library was heterozygous for both proviruses. Since most human allelic variation probably arose within the last 1,000,000 years [6–9], both HERV-K113 and HERV-K115 likely formed within this time period.

PCR products were amplified only from orthologous loci

We considered the possibility that the PCR primers used to detect the preintegration sites might amplify sequences from other loci in the human genome and thus yield erroneous data about which alleles were present. RepeatMasker and BLAST analyses showed that HERV-K113 was inserted in a 785 bp stretch of human DNA sequence between a SINE/AluY element and a LINE1/MA2 element that was part of a large, low-copy repeat sequence. Most of the low-copy repeats, like the one containing HERV-K113, were located on chromosome 19. Comparison of 500 bp immediately flanking the provirus insertion site to the corresponding stretches in the other repeats showed that all were less than 90% identical to the sequences flanking the provirus. In contrast, the orthologous preintegration site on BAC RP11-678G14 had only one nucleotide difference from the sequences flanking the provirus over the same 500 bp stretch. HERV-K115 was inserted at a junction between ancient SINE/MIR and LINE/CR1 (L3) elements. No other sequences in GenBank showed significant similarity to those flanking HERV-K115, except for the three BACs containing the preintegration site. The four sequences differed from each other at three positions over the 500 bp immediately flanking the provirus, and these probably represent single nucleotide polymorphisms among humans. In summary, each proviral locus has a sufficiently unique sequence to be distinguished easily from any other locus in the human genome. The sequences of the preintegration sites in GenBank matched those flanking the corresponding proviruses and differed substantially from those of any other loci in the human genome.

For each provirus, the proviral junction and preintegration site PCR fragments from several individuals (Figure 1c) were sequenced and found to match precisely the sequences for the proviral and preintegration site alleles. Thus, the PCR products were amplified strictly from the orthologous loci. In particular, the preintegration site PCR products (A + D, Figure 1c) for HERV-K113 were not derived from any other low-copy repeat on chromosome 19.

Figure 2



Genetic structures of HERV-K113 and HERV-K115 proviruses. Positions of the viral open reading frames (ORFs) including ATG start codons and stop codons are shown. *c* indicates the first coding exon of the cORF/KRev protein. SD and SA indicate splice donors and splice acceptors. The triangle below the HERV-K115 genome indicates the position of the 1 bp deletion relative to other full-length HERV-K proviruses. The boxes below the map of each provirus depict the three possible translational reading frames with −1 shifts from any row to the one below it. In HERV-K115, the 1 bp deletion near the *gag-pro* boundary resulted in replacement of the normal 31 amino acids at the carboxyl terminus of the Gag precursor protein with 12 amino acids from what is the +1 ORF in other HERV-Ks. The novel sequences are shown as the very small black box at the C terminus of the *gag* ORF. Even though the *pro* and *pol* ORFs of HERV-K115 are present, they are unlikely to be translated due to the 1 bp deletion near the *gag-pro* boundary and are thus excluded from the figure.

Preintegration site alleles were not generated by deletion of proviruses by recombination

The sequences of the preintegration site PCR products (A + D, Figure 1c) also argued strongly against the possibility that the preintegration site alleles were generated by replacement of an existing provirus by a recombination event such as gene conversion involving a nonorthologous locus containing sequences similar to those flanking the provirus. If such an event had occurred, sequences for at least a short distance on either side of the provirus would also have been replaced, and the preintegration site sequence would be expected to differ from the sequences flanking the corresponding provirus. However, this was not the case. The sequences of the amplified products (A + D, Figure 1c) always matched those of the sequences flanking the proviruses. Thus, it is highly unlikely that the preintegration site allele for either provirus actually resulted from proviral loss due to a recombination event. In summary, the PCR analyses (Figure 1c) robustly determined whether the authentic preintegration site and proviral alleles were present in human DNA samples.

Proteins encoded by the proviruses

To determine the coding capacity of each provirus, the viral genomes were sequenced (Figure 2). HERV-K113 was found to have full-length open reading frames (ORFs) for all viral proteins, with no substitutions that would alter amino acid sequence motifs that are well conserved among retroviruses. This distinguishes HERV-K113 from HERV-K(HLM-2.HOM) [also called HERV-K108 and HERV-K(C7)], which encodes CIDD instead of a standard YIDD motif in reverse transcriptase, and from other previously described HERV-Ks [1, 3, 4, 10]. Thus, HERV-K113 is the first endogenous retrovirus described in humans with full-length ORFs for all viral proteins and no amino acid substitutions in conserved sequence motifs.

Relative to other HERV-K proviruses [1], HERV-K115 has a 1 bp deletion located 92 bp upstream from the stop codon of the *gag* ORF (Figure 2). This mutation alters the carboxyl terminus of the encoded Gag precursor protein and alters the ribosomal frameshift required to translate the *pro* and *pol* ORFs [11] from the standard −1 of the *Betaretroviruses*, the genus to which HERV-K belongs along with the related mouse mammary tumor virus and type-D primate viruses, to +1. Thus, it is unlikely that the *pro* and *pol* ORFs can be translated from HERV-K115. HERV-K115 does encode a full-length cORF/KRev protein that functions in the nuclear export of viral RNA [12, 13]. It also encodes a full-length Env protein. The evolutionary conservation of full-length *env* ORFs in at least four HERV-K proviruses, HERV-Ks 115, 113, 109, and 108/(HLM-2.HOM)/(C7) (Figure 2 and [1]), strongly suggests that the HERV-K Env protein is crucial for viral reinfection of germ cells. This implies that, at least much of the time, such reinfections involve standard retroviral particles containing functional HERV-K Env protein that enter the cells via a cellular receptor for the virus.

Estimation of proviral ages by LTR sequence comparisons

The relative ages of endogenous retroviruses can be estimated by comparing the sequences of the two viral long-terminal repeats (LTRs) [14–18]. Due to the mechanism of reverse transcription, retroviral LTRs are usually identical at the time when a provirus forms. Mutations that subsequently accumulate over evolutionary time are unique to one of the two LTRs. Since both HERV-K113 and HERV-K115 are full-length proviruses, LTR sequences can be used to estimate how long ago they formed. Besides these two proviruses, there is a HERV-K solo LTR that is in the HLA-DQB1 locus of some but not all humans [2, 19]. However, since it is not a full-length provirus with two LTRs, this type of analysis cannot be applied to estimate its age.

The two LTRs of HERV-K115 had 14 differences: 13 single bp substitutions and an 8 bp deletion at 115 bp

from the beginning of the 3′ LTR (TCTGTTAATC TATGACCT, deleted bases underlined). Nine of these mutations, including the 8 bp deletion, were previously observed in both LTRs of multiple other HERV-K proviruses in humans [1]. Thus, it is highly unlikely that these represent de novo mutations in HERV-K115. The most likely explanation for them is that they arose by a recombination between two HERV-K genomes, perhaps a gene conversion event involving a second HERV-K locus in the human genome and one of the HERV-K115 LTRs. Gene conversion of endogenous retroviral loci was previously documented to have occurred [18]. Thus, at most only five of the differences between the HERV-K115 LTRs likely arose by mutation over evolutionary time, and even some of those might have been introduced by the same gene conversion event. Thus, estimation of the relative age of HERV-K115 based on LTR sequence differences is potentially inaccurate, although a total of five differences probably provides an upper limit for estimating how long ago it formed.

There were no differences between the two LTRs of HERV-K113. This is consistent with HERV-K113 being a very new provirus. By comparing the number of differences between LTRs of individual HERVs and the dates of divergence of the most distantly related species containing them, it was estimated that endogenous retroviruses accumulate mutations at a rate of roughly 2.3×10^{-9} to 5×10^{-9} substitutions per site per year [18]. That equals one difference per LTR every 200,000–450,000 years (HERV-K LTRs are about 970 bp long). Similarly, it was estimated based on intronic mutation rates that HERV-K(HLM-2.HOM), which has accumulated six differences between its LTRs, formed about 1.2 million years ago [3]. Using these values, it can be estimated that HERV-K113, which has no differences between its LTRs, likely formed more recently than 200,000–450,000 years ago and perhaps considerably more recently.

Implications for HERV-K infectivity

HERV-K first infected the germline of the lineage leading to humans roughly 35 million years ago, sometime after the divergence of platyrrhines (New World monkeys) from catarrhines (cercopithecoids and hominoids), but before the separation of cercopithecoids (Old World monkeys) from hominoids (apes and humans). More distantly related sequences have been described for platyrrhines [20]. The existence of insertional polymorphisms of HERV-K113 and HERV-K115 in humans, the absence or low number of nucleotide differences between the LTRs of the proviruses, and the presence of full-length ORFs for all viral proteins in HERV-K113 support the idea that these two proviruses were very recent additions to the human genome. Thus, HERV-K reinfected humans in very recent evolutionary times. Neighbor-joining and maximum-likelihood analyses showed that both HERV-K113 and HERV-K115 are very closely related to other human-specific HERV-K proviruses. Thus, they descended from earlier human HERV-Ks and were not the result of recent crossspecies transmission. Each virus was detected in at least one sub-Saharan African and in at least one non-African. This is consistent with the proviruses having originally formed in two individuals in Africa prior to the emergence of modern humans from Africa, but recently enough that neither proviral allele was fixed in the relatively small human population at that time. Evidence suggests that the emergence of modern humans from Africa began perhaps 100,000 years ago [21, 22].

Analysis of the human genome sequence led to the idea that replication-competent endogenous retroviruses in the human genome may be extinct or very nearly so [23]. However, neither HERV-K113 nor HERV-K115 was previously included in GenBank. Moreover, since both HERV-K113 and HERV-K115 were initially identified by screening just a single individual (RP11), it is reasonable to hypothesize that there may be additional, more recently acquired HERV-K proviruses that are present at even lower frequencies among humans today. Strong data support the conclusion that HERV-K has been infectious in the human lineage from about 35 million years ago through the time of formation of HERV-K113 which may have occurred as recently as within the last 100,000 years. Unless the virus has suddenly lost its ability to replicate and reinfect the human germline after being active for about 35 million years, HERV-K should still be infectious in humans today. Recent HERV-K infections might have occurred by complementation involving viral proteins encoded by different HERV-K proviruses in the human genome, or they may have involved individual proviruses. HERV-K113 is the best candidate to be a single provirus that is active in humans today.

Materials and methods

BAC screening and PCR

BAC library screening, PCR reactions with Taq polymerase for products <1 kb, and PCR with Expand Long Template PCR System (Boehringer-Mannheim) for products >1 kb were performed as described [1]. Amplification of genomic sequences flanking proviruses in BACs by inverse PCR was performed as described [1, 4, 24].

Supplementary material

Supplementary material including primer sequences, GenBank accession numbers, BAC addresses, and methods of DNA sequence analysis is available at http://images.cellpress.com/supmat/supmatin.htm.

Acknowledgements

We thank R. Kim for helpful discussions. This work was supported by research grant CA44822 and training grant GM07491 from the National Institutes of Health and by grant BC996431 from the United States Army Medical Research and Materials Command.

References

1. Barbulescu M, Turner G, Seaman MI, Deinard AS, Kidd KK, Lenz J: **Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans.** *Curr Biol* 1999, **9**:861-868.
2. Medstrand P, Mager DL: **Human-specific integrations of the**

HERV-K endogenous retrovirus family. *J Virol* 1998, **72:**9782-9787.
3. Mayer J, Sauter M, Racz A, Scherer D, Mueller-Lantzsch N, Meese E: **An almost-intact human endogenous retrovirus K on human chromosome 7.** *Nat Genet* 1999, **21:**257-258.
4. Tönjes RR, Czauderna F, Kurth R: **Genome-wide screening, cloning, chromosomal assignment, and expression of full-length human endogenous retrovirus type K.** *J Virol* 1999, **73:**9187-9195.
5. Brown PO: **Integration.** In *Retroviruses*. Edited by Coffin JM, Hughes SH, Varmus HE. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997:343-435.
6. Takahata N: **Allelic genealogy and human evolution.** *Mol Biol Evol* 1993, **10:**2-22.
7. Zhao Z, Jin L, Fu YX, Ramsay M, Jenkins T, Leskinen E, *et al.*: **Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22.** *Proc Natl Acad Sci USA* 2000, **97:**11354-11358.
8. Alonso S, Armour JA: **A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa.** *Proc Natl Acad Sci USA* 2000, **3:**864-869.
9. Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, *et al.*: **Archaic African and Asian lineages in the genetic ancestry of modern humans.** *Am J Hum Genet* 1997, **60:**772-789.
10. Bock M, Stoye JP: **Endogenous retroviruses and the human germline.** *Curr Opin Genet Dev* 2000, **10:**651-655.
11. Swanstrom R, Wills JW: **Synthesis, assembly, and processing of viral proteins.** In *Retroviruses*. Edited by Coffin JM, Hughes SH, Varmus HE. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997:263-334.
12. Magin C, Löwer R, Löwer J: **cORF and RcRE, the Rev/Rex and RRE/RxRE homologues of the human endogenous retrovirus family HTDV/HERV-K.** *J Virol* 1999, **73:**9496-9507.
13. Yang J, Bogerd HP, Peng S, Wiegand H, Truant R, Cullen BR: **An ancient family of human endogenous retroviruses encodes a functional homolog of the HIV-1 Rev protein.** *Proc Natl Acad Sci USA* 1999, **96:**13404-13408.
14. Shih A, Coutavas EE, Rush MG: **Evolutionary implications of primate endogenous retroviruses.** *Virology* 1991, **182:**495-502.
15. Dangel AW, Baker BJ, Mendoza AR, Yu CY: **Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution.** *Immunogenetics* 1995, **42:**41-52.
16. Mager DL, Freeman JD: **HERV-H endogenous retroviruses: presence in the New World branch but amplification in the Old World primate lineage.** *Virology* 1995, **213:**395-404.
17. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: **The paleontology of intergene retrotransposons of maize.** *Nat Genet* 1998, **20:**43-45.
18. Johnson WE, Coffin JM: **Constructing primate phylogenies from ancient retrovirus sequences.** *Proc Natl Acad Sci USA* 1999, **96:**10254-10260.
19. Donner H, Tönjes RR, Bontrop RE, Kurth R, Usadel KH, Badenhoop K: **Intronic sequence motifs of HLA-DQB1 are shared between humans, apes and Old World monkeys, but a retroviral LTR element (DQLTR3) is human specific.** *Tissue Antigens* 1999, **53:**551-558.
20. Simpson GR, Patience C, Löwer R, Tönjes RR, Moore HD, Weiss RA, *et al.*: **Endogenous D-type (HERV-K) related sequences are packaged into retroviral particles in the placenta and possess open reading frames for reverse transcriptase.** *Virology* 1996, **222:**451-456.
21. Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, *et al.*: **Global patterns of linkage disequilibrium at the CD4 locus and modern human origins.** *Science* 1996, **271:**1380-1387.
22. Klein RG: *The Human Career: Human Biological and Cultural Origins*. Chicago: University of Chicago Press; 1999.
23. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
24. Li J, Shen H, Himmel KL, Dupuy AJ, Largaespada DA, Nakamura T, *et al.*: **Leukaemia disease genes: large-scale cloning and pathway predictions.** *Nat Genet* 1999, **23:**348-353.