

A HERV-K provirus in chimpanzees, bonobos and gorillas, but not humans

Madalina Barbulescu*[§], Geoffrey Turner*[§], Mei Su*, Rachel Kim*, Michael I. Jensen-Seaman^{†||}, Amos S. Deinard^{†##}, Kenneth K. Kidd[†] and Jack Lenz*

Evidence from DNA sequencing studies strongly indicated that humans and chimpanzees are more closely related to each other than either is to gorillas [1–4]. However, precise details of the nature of the evolutionary separation of the lineage leading to humans from those leading to the African great apes have remained uncertain. The unique insertion sites of endogenous retroviruses, like those of other transposable genetic elements, should be useful for resolving phylogenetic relationships among closely related species. We identified a human endogenous retrovirus K (HERV-K) provirus that is present at the orthologous position in the gorilla and chimpanzee genomes, but not in the human genome. Humans contain an intact preintegration site at this locus. These observations provide very strong evidence that, for some fraction of the genome, chimpanzees, bonobos, and gorillas are more closely related to each other than they are to humans. They also show that HERV-K replicated as a virus and reinfected the germline of the common ancestor of the four modern species during the period of time when the lineages were separating and demonstrate the utility of using HERV-K to trace human evolution.

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

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

[§]These authors contributed equally to this manuscript.

Present addresses: ^{||}Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA. [#]School of Veterinary Medicine, University of California, Davis, Davis, California 95616, USA.

Received: 21 August 2000
Revised: 22 February 2001
Accepted: 28 February 2001

Published: 15 May 2001

Current Biology 2001, 11:779–783

0960-9822/01/\$ – see front matter
© 2001 Elsevier Science Ltd. All rights reserved.

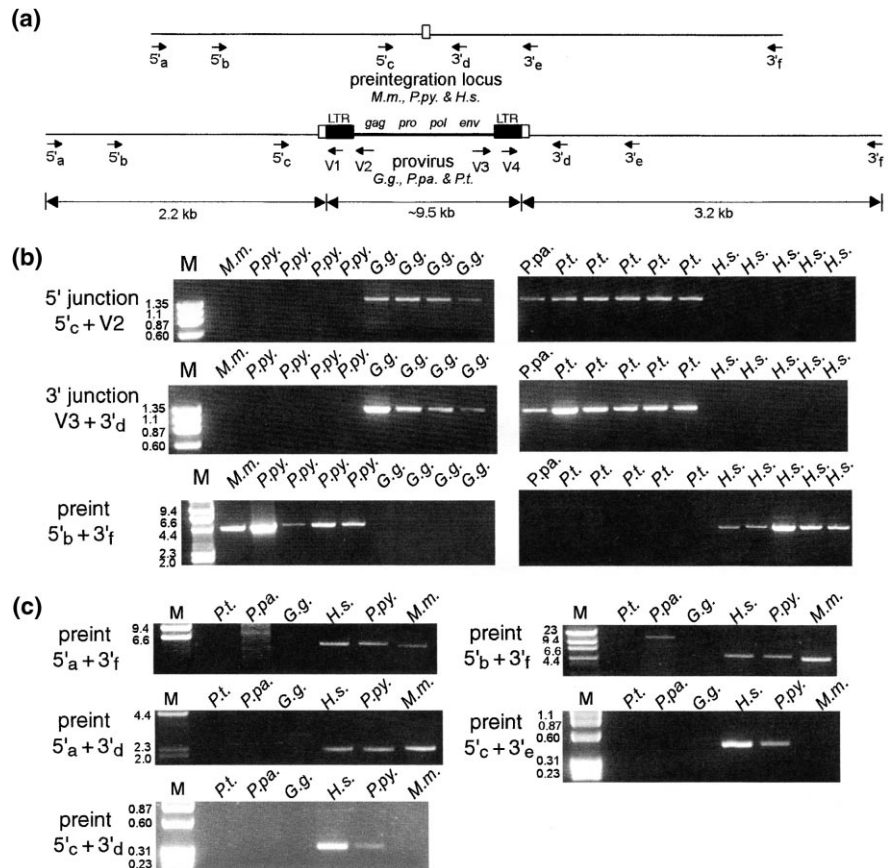
Results and discussion

Endogenous retroviruses exist in distinct structural forms that proceed in a defined chronological order. As with other retrotransposable elements, the preintegration site is the most ancestral allele [5–7] and is followed by the inserted element, which is the full-length provirus in the case of endogenous retroviruses. Frequently, homologous recombination between the two long terminal repeats (LTRs) of full-length proviruses then results in the formation of solo LTRs. Random deletions of parts of proviruses along with flanking host sequences may also occur. Proviruses or solo LTRs present at the same site in the genomes of two species are identical by descent, as the likelihood of independent integrations at the same site (insertional homoplasy) is negligible [7, 8].

HERV-K [9] is present in the genomes of catarrhines (cercopithecoids = Old World monkeys, and hominoids = apes and humans) [10–12], and more distantly related sequences are present in platyrrhines (New World monkeys) [13]. It has been reinfected the germline of the lineage leading to modern humans in recent evolutionary time [14, 15]. Many of the HERV-K proviruses present in the human genome today formed after the evolutionary separation of the human lineage from the chimpanzee and gorilla lineages [14, 15]. Others formed prior to the separation of the three genera and are present at orthologous positions in the human, chimpanzee, bonobo, and gorilla genomes, but not in the orangutan genome [14]. Therefore, HERV-K was active both before and after the evolutionary separation of humans (*Homo sapiens*), common chimpanzees (*Pan troglodytes*), bonobos (pygmy chimpanzees, *Pan paniscus*), and gorillas (*Gorilla gorilla*) from a common ancestor. If it was also active during the period when the lineages leading to the modern species were separating, then the insertion sites of HERV-K proviruses could be useful for tracing those lineages. To date, no sites of HERV-K provirus insertion, or those of any mobile genetic element, have been reported to be in only two of the three genera. To examine the evolutionary history of HERV-K among hominoids, we identified HERV-K proviruses in the gorilla genome by inverse PCR amplification of the 5' provirus-host junctions using conditions described by others [16, 17] and primers described previously [14]. The DNA sequences flanking the proviruses were determined. A BLAST search of the human genome using the sequence flanking one such provirus (HERV-K-GC1) indicated that the orthologous position

Figure 1

PCR amplification of HERV-K-GC1 from primate genomic DNAs. Species abbreviations are: *M.m.*, *Macaca mulatta*; *P.py.*, *Pongo pygmaeus*; *G.g.*, *Gorilla gorilla*; *P.pa.*, *Pan paniscus*; *P.t.*, *Pan troglodytes*; and *H.s.*, *Homo sapiens*. **(a)** A map of the HERV-K-GC1 preintegration and proviral loci. Arrows show the positions of PCR primers used in this study. Primers within HERV-K were from Barbulescu et al. [14]. Those from the flanking sequences were based on the human sequence from BAC RPC11-500M8 (GenBank Accession number AC005832). *Macaca mulatta* lacks a stretch of about 600 bp relative to the hominoid species near primer 3'f. The white boxes indicate the 5 bp target sequence that was duplicated during the integration of HERV-K-GC1. The black rectangles denote the HERV-K long terminal repeats (LTRs). **(b)** PCR amplification of the provirus junctions and preintegration site. Each lane shows the products from a different individual of the indicated species. M, marker. **(c)** PCR amplification of the HERV-K-GC1 provirus and preintegration site using different primer pairs in one individual of each of the indicated species.



in humans lacked a HERV-K provirus or solo LTR. PCR primers were designed based on the human sequence (BAC RP11-500M8 from human chromosome 12p13.3, GenBank accession number AC005832) that flanked both sides of the HERV-K-GC1 provirus and used to amplify the orthologous sequences from various primates (Figure 1). Sequencing of the resulting products (Figure 2) revealed that integration of HERV-K-GC1 involved a 5 bp target site duplication (5'-ATTAT-3' flanking the viral + strand) and that the provirus was inserted at the identical bp in the genomes of gorillas, bonobos, and common chimpanzees. No preintegration site allele was detected by PCR in any *Gorilla* or *Pan* individual tested (Figure 1). Thus, the HERV-K-GC1 provirus must have formed prior to the separation of the lineages leading to *Gorilla* and *Pan*.

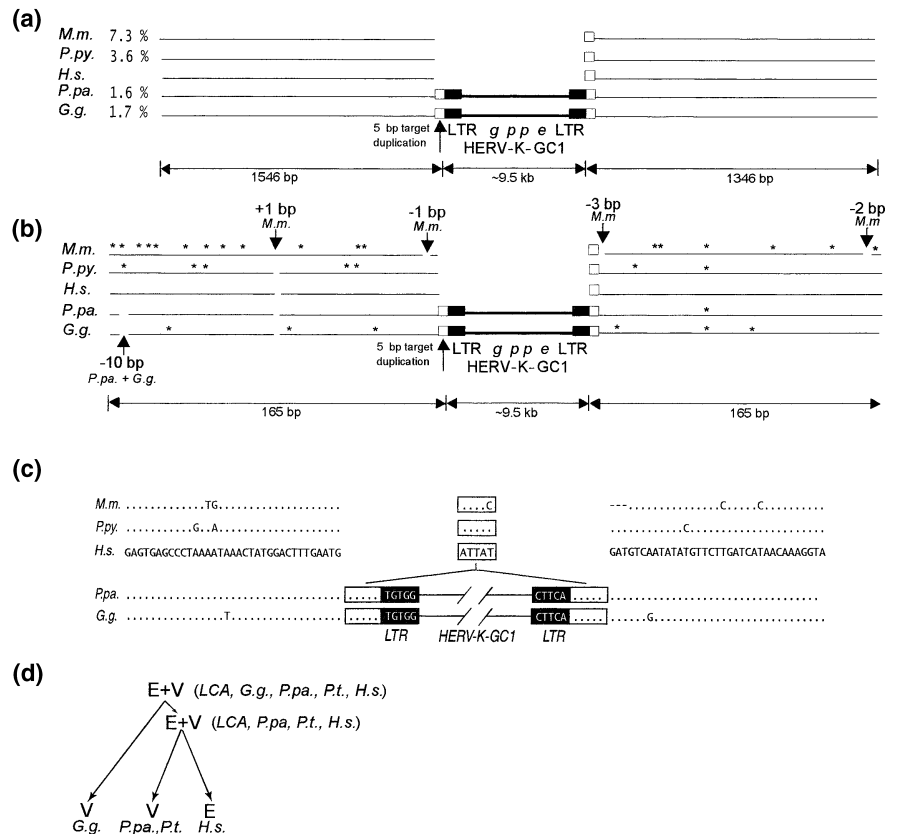
In contrast, PCR and sequencing analyses showed that the human, orangutan (*Pongo pygmaeus*), and rhesus macaque (*Macaca mulatta*) genomes contained a preintegration site at the orthologous locus (Figures 1 and 2). No solo LTR was present, and the 5 bp target site was not duplicated. Multiple humans and orangutans were tested, and all were found to contain only the preintegration site (Figure 1b).

Several possibilities were considered to explain how a provirus could be present in *Gorilla* and *Pan* but be absent in *Homo*. It is highly unlikely that the provirus was deleted in humans, as the retroviral integration process is irreversible. Another possibility was that the provirus was replaced in the human lineage by a gene conversion or unequal crossover event. In particular, the preintegration site may have been duplicated either in tandem or at another position within the genome of the common ancestor of *Homo*, *Pan*, and *Gorilla*. A recombination event involving the duplicated locus could then have replaced the 9.5 kb provirus in humans with a sequence similar to the preintegration site. In this regard, analysis of the human sequence flanking the HERV-K-GC1 integration site in *Pan* and *Gorilla* indicated that the ape provirus lies within an older L1 retrotransposon and that several L1 elements and an Alu element lie within a 5 kb stretch flanking the insertion site of the provirus. This particularly raised the possibility that gene conversion from an L1 element at a nonorthologous position might have replaced the provirus in the human lineage.

To test these possibilities, we designed several PCR primers based on the human sequence over a 5.4 kb stretch

Figure 2

Sequences flanking the HERV-K-GC1 insertion site in different primate species. Sequences were entered into GenBank (Accession numbers AF294259-AF294264). **(a)** A diagram showing the sequenced region relative to the position of the provirus in *Gorilla* and *Pan*. The numbers to the left of each line show the differences between the corresponding sequence and the human sequence, in which each substitution, insertion, or deletion relative to human was counted as one difference. The black rectangles indicate the HERV-K long terminal repeats (LTRs). The white boxes indicate the 5 bp (ATTAT) that were duplicated flanking the provirus in *Gorilla* and *Pan*. These are present only once in *Homo*, *Pongo*, and *Macaca*. The 5 bp sequence is present at nucleotide 173,033 of version AC005832.1 of the BAC. GENSCAN 1.0 analysis of the human BAC predicted the presence of a gene similar to one of unknown function predicted within the *Drosophila melanogaster* genome (CG9986, GenBank Accession numbers AAF56802) ~2 kb downstream of the position of the provirus in *Gorilla* and *Pan* and in the opposite transcriptional orientation. RepeatMasker analysis of the human BAC indicated that there were several L1 sequences and an Alu element within a 5 kbp stretch surrounding the provirus insertion site in apes. **(b)** The sequences corresponding to 165 bp flanking either side of the provirus. Asterisks indicate the positions of single base pair substitutions. Deletions in the indicated species are shown as gaps. **(c)** Sequences of 40 bp immediately flanking either side of the provirus. Dots indicate sequence identity relative to the human



sequence. Dashes indicate the absence of the corresponding nucleotide. The last five nucleotides of the viral LTRs are shown in the black rectangles. **(d)** Segregation of the empty preintegration allele (E) and the

provirus allele (V) in the *Homo*, *Pan*, and *Gorilla* lineages. E + V indicates that both alleles were present in the population of the cognate species. LCA, last common ancestor.

of sequence flanking the site of the HERV-K-GC1 insertion in *Pan* and *Gorilla* (Figure 1). No evidence for a preintegration locus in *Pan* or *Gorilla* was seen with any combination of primers used (Figure 1c). Thus, neither of those genera contains a duplicated locus, tandem or otherwise, including any L1 element at a nonorthologous position, that is sufficiently similar to the HERV-K-GC1 site to be recognized by any of the PCR primer pairs used. The data are consistent with the conclusion that these genera lack an appropriate locus for a putative gene conversion event that could have eliminated the provirus within the human lineage.

We also considered the possibility that a putative recombination event involved a duplication of a sequence flanking the provirus insertion site that was too short to be detected with the PCR primers used. The innermost primer pair used (5'c and 3'd, Figure 1) generated a 299 bp product from human DNA containing 192 bp 5' to the ATTAT duplication in apes and 102 bp 3' to the duplication. To look for evidence of such a short, tandem duplication, the

sequences flanking the provirus in a gorilla and a bonobo and the corresponding preintegration loci from an orangutan and a rhesus macaque were determined by the sequencing of PCR-amplified segments of the cognate genomes and compared to the corresponding stretch of human DNA (Figure 2). A stretch corresponding to 2852 bp in humans was determined in each species. The five sequences were colinear throughout their entire lengths, except for the provirus and indels of a few bp, such as those shown in Figure 2b. Thus, there is no small duplication, tandem or otherwise, within the sequenced stretch of any of the genera that might have participated in a putative recombination event to replace the provirus within the human lineage. The colinearity also showed that all the L1 elements within the sequenced stretch flanking the proviral insertion site were ancestral to the cercopithecoid-hominoid divergence. As expected, the *Homo* sequence was most related to those of the African apes, *Pan* and *Gorilla*, and more distantly related to those of *Pongo* and *Macaca* (Figure 2a). In pairwise comparisons, there was a 1.5% difference between *Pan* and *Gorilla*, 1.6%

difference between *Homo* and *Pan*, and a 1.7% difference between *Homo* and *Gorilla*. Positions where two of the three species possessed a shared derived difference relative to the outgroup sequence provided by *Pongo* and *Macaca* provided insight into relationships among genera. *Homo* and *Gorilla* shared one such difference (a single bp substitution), while *Pan* and *Homo* shared none. However, *Pan* and *Gorilla* shared seven differences, including four single bp substitutions, a one bp deletion, a ten bp deletion 160 bp 5' to the provirus (Figure 2b), and the provirus itself. These observations strongly supported *Pan* and *Gorilla* as being the most closely related genera in this part of the genome.

The similarity of the *Homo* sequence to those of *Pan* and *Gorilla* is clear, as it differs by 1.6%–1.7% (Figure 2). BLAST searches of the human genome (htgs and nr databases) with sequences as small as 45 bp immediately flanking the provirus (ATTAT target for HERV-K-GC1 plus 20 bp on either side) showed that the next best matches after the orthologous locus differed by 8%–10% from the sequences immediately flanking the HERV-K-GC1 provirus in *Pan* and *Gorilla*. This indicates that the L1 element into which the HERV-K-GC1 provirus is integrated is an ancient element that has accumulated a sufficient number of unique mutations to be 8%–10% different from any other L1 element in the human genome. Thus, it is unlikely that any nonorthologous sequence in the human genome, L1 repeat or otherwise, existed in recent human evolution that could have served as the source sequence for a putative gene conversion event that replaced the HERV-K-GC1 provirus specifically within the human lineage, even an event that replaced as little as 20 bp on either side of the 9.5 kbp HERV-K-GC1 provirus. Rather, the human locus is clearly more closely related to the orthologous loci in *Pan* and *Gorilla* than it is to any other locus in the human genome.

In principle, the following gene conversion scenario could have resulted in the removal of the provirus in the human lineage. First, the preintegration locus underwent a duplication event in the common ancestor of *Homo*, *Pan*, and *Gorilla*. Second, the provirus formed in one of the two copies of the locus by viral infection of the common ancestor. Next, the *Gorilla* lineage diverged from the *Pan-Homo* common ancestor. Then, the *Pan* and *Homo* lineages diverged. Afterwards, a recombination event reversed the original locus duplication, restoring a single copy of the locus without the provirus in the *Homo* lineage. However, the PCR and sequencing assays uniformly failed to detect any evidence for the presence of such a duplicated locus in *Gorilla* or *Pan*. Therefore, in addition to the removal of the provirus specifically in the *Homo* lineage, the scenario also requires recombination events in the *Pan* and *Gorilla* lineages to eliminate the provirus-free copy of the

locus. Since the *Gorilla* lineage diverged before the *Pan* and *Homo* lineages separated, this means that independent recombination events would have had to occur in both the *Pan* and *Gorilla* lineages. While this scenario can never be formally excluded, there is a more parsimonious alternative that involves three fewer recombination events.

The alternative is an allelic segregation model (Figure 2d) in which the provirus formed in the most recent common ancestor of *Homo*, *Pan*, and *Gorilla* just before the three lineages separated. The provirus allele was fixed in the *Gorilla* lineage. Both alleles were then maintained in the *Pan-Homo* common ancestor until the individual lineages diverged. The provirus allele was fixed in the *Pan* lineage, while the preintegration site allele was fixed in the *Homo* lineage. The allelic segregation model is more parsimonious than the gene conversion scenario because it does not require the locus duplication event in the common ancestor or the two independent losses of the duplicated locus in the *Gorilla* and *Pan* lineages. In addition, the possibility that humans diverged first, the provirus formed next, and the gorilla-chimpanzee divergence occurred last is extremely unlikely given the greatest sequence similarity between chimpanzees and humans at most loci [1–4]. Rather, the presence of HERV-K-GC1 in gorillas and chimpanzees, but not humans, is best explained by the maintenance of the preintegration site in the human lineage since before the time when the provirus formed in the common ancestor of chimpanzees and gorillas. This leads to the conclusion that, for some fraction of the genome, the gorilla and chimpanzee genomes are more closely related to each other than either is to humans. Furthermore, the number of shared derived differences linking chimpanzees with gorillas (seven in 2852 bp) suggests that relatively divergent haplotypes existed in the last common ancestor of the three modern genera, which is indicative of a large effective population size. This is similar to the pattern observed in extant great apes, in which haplotypes differing by up to nine substitutions per kb are found within populations of chimpanzees or gorillas [18, 19] and are consistent with recent estimates of the effective population size of the last common ancestor of humans and chimpanzees [4].

The precise details of the nature of the phylogenetic separation of humans from the African great apes has remained uncertain. Genetic studies indicated that humans and chimpanzees are the most closely related pair for much of the genome [1–4]. However, for some fraction of the genome, they are not [1, 3, 4]. Such data are consistent with a model in which alleles segregated differently among the three eventual lineages [1, 4, 19–21]. Some alleles that were polymorphic in the common ancestor of gorilla, chimpanzees, and humans became fixed within the common ancestor of humans and chimpanzees before

the latter two lineages separated. This, along with new mutations in the human-chimpanzee common ancestor, accounts for the higher genetic relatedness of chimpanzees and humans that appears to encompass the majority of the genome [1, 3, 4]. However, at positions in the genome where allelism was maintained throughout the period of existence of the human-chimpanzee common ancestor, some of the same alleles that became fixed in the gorilla lineage may also have been fixed in only one of the human or chimpanzee lineages. The HERV-K-GC1 provirus provides a compelling piece of evidence for such a model, as it is the clearest example to date of a specific locus within the genome where chimpanzees and gorillas are more closely related to each other than either is to humans. Moreover, since neutral alleles are maintained in a population for only a limited time that depends on the size of the population [4, 19–24], the data presented here imply that the separation of the *Homo*, *Pan*, and *Gorilla* lineages occurred within a period of time that was sufficiently short for such allelism to be maintained. The significance of the work presented here is the demonstration of the utility of HERV-K as a marker for studying human evolution, the conclusion that HERV-K was active at about the time that the three lineages were evolutionarily separating, and the very strong experimental evidence that, in some fraction of the genome, chimpanzees, bonobos, and gorillas are more closely related to each other than any of them is to humans. HERV-K and other retrotransposable elements should contribute to determining what that fraction is.

Acknowledgements

Ape tissue samples were kindly provided by Stanford University, the Yerkes Regional Primate Center, the Henry Doorly Zoo, and the Milwaukee County Zoo. This work was supported by research grant CA44822 and training grants GM07491 and CA09060 from the National Institutes of Health.

References

- Satta Y, Klein J, Takahata N: **DNA archives and our nearest relative: the trichotomy problem revisited.** *Mol Phylogenet Evol* 2000, **14**:259-275.
- Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, et al.: **Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence.** *Mol Phylogenet Evol* 1998, **9**:585-598.
- Ruvolo M: **Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets.** *Mol Biol Evol* 1997, **14**:248-265.
- Chen FC, Li WH: **Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees.** *Amer J Hum Gen* 2001, **68**:444-456.
- Stoneking M, Fontius JJ, Clifford SL, Soodyall H, Arcot SS, Saha N, et al.: **Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa.** *Genome Res* 1997, **7**:1061-1071.
- Hamdi H, Nishio H, Zielinski R, Dugaiczky A: **Origin and phylogenetic distribution of Alu DNA repeats: irreversible events in the evolution of primates.** *J Mol Biol* 1999, **289**:861-871.
- Boeke JD, Stoye JP: **Retrotransposons, endogenous retroviruses, and the evolution of retroelements.** In *Retroviruses*. Edited by Coffin JM, Hughes SH, and Varmus HE. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1997:343-435.
- Withers-Ward ES, Kitamura Y, Barnes JP, Coffin JM: **Distribution of targets for avian retrovirus DNA integration in vivo.** *Genes Dev* 1994, **8**:1473-1487.
- Ono M: **Molecular cloning and long terminal repeat sequences of human endogenous retrovirus genes related to types A and B retrovirus genes.** *J Virol* 1986, **58**:937-944.
- Mariani-Costantini R, Horn TM, Callahan R: **Ancestry of a human endogenous retrovirus family.** *J Virol* 1989, **63**:4982-4985.
- Steinhuber S, Brack M, Hunsmann G, Schwelberger H, Dierich MP, Vogetseder W: **Distribution of human endogenous retrovirus HERV-K genomes in humans and different primates.** *Hum Genet* 1995, **96**:188-192.
- Dangel AW, Mendoza AR, Baker BJ, Daniel CM, Carroll MC, Wu LC, et al.: **The dichotomous size variation of human complement C4 genes is mediated by a novel family of endogenous retroviruses, which also establishes species-specific genomic patterns among Old World primates.** *Immunogenetics* 1994, **40**:425-436.
- Simpson GR, Patience C, Lower R, Tonjes 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.
- 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.
- Medstrand P, Mager DL: **Human-specific integrations of the HERV-K endogenous retrovirus family.** *J Virol* 1998, **72**:9782-9787.
- 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.
- Tonjes 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.
- Deinard AS, Kidd KK: **Identifying conservation units within captive chimpanzee populations.** *Amer J Phys Anthropol* 2000, **111**:25-44.
- Deinard A, Kidd K: **Evolution of a HOXB6 intergenic region within the great apes and humans.** *J Hum Evol* 1999, **36**:687-703.
- Pamilo P, Nei M: **Relationships between gene trees and species trees.** *Mol Biol Evol* 1998, **5**:568-583.
- Wu CI: **Inferences of species phylogeny in relation to segregation of ancient polymorphisms.** *Genetics* 1991, **127**:429-435.
- Takahata N: **Allelic genealogy and human evolution.** *Mol Biol Evol* 1993, **10**:2-22.
- Kaessmann H, Wiebe V, Pääbo S: **Extensive nuclear DNA sequence diversity among chimpanzees.** *Science* 1999, **286**:1159-1162.
- Garner KJ, Ryder OA: **Mitochondrial DNA diversity in gorillas.** *Mol Phylogenet Evol* 1996, **6**:39-48.