The structure and phylogeny of a new family of human endogenous retroviruses

Bengt Widegren,1 Christian Kjellman,1 Stefan Aminoff,1 Leif G. Sahlford2 and Hans-Olov Sjögren1

1 Department of Tumor Immunology, Institute of Cell and Molecular Biology, The Wallenberg Laboratory, Solvegatan 33, S-223 62 Lund, Sweden
2 Department of Neurosurgery, The Academy Hospital, S-221 85 Lund, Sweden

A novel endogenous retrovirus (ERV) designated XA34 was isolated from a human glioma cDNA library using low stringency hybridization with an ERV-9 env probe. Southern blot hybridizations with human genomic DNA revealed the presence of approximately 16 genomic copies closely related to XA34. Sequencing of a 2303 bp cDNA clone of XA34 showed that it belongs to a new ERV family. The XA34 ERV has recombined with an ERV-9-1ike retrovirus resulting in a truncated ERV-9-1ike env region that ends with an Alul-like 3' LTR. By using PCR, we isolated ~ 940 bp polfragments from three additional members of this family, XA35, XA36 and XA37. A fifth member, XA38, was isolated and sequenced as a 4729 bp genomic clone. The genomic XA38 clone spans from pol towards the 3' flanking region. The XA38 virus contains a more cryptic env region. The XA38 env is truncated in the transmembrane region and the virus then ends with three Alu repeats. Southern blot studies with human, chimpanzee, orangutan and squirrel monkey DNA show the presence of the XA34 family in all these species. That both the New and Old World monkeys have this ERV family means that the integration and/or amplification in the primate germ-line of XA34 probably took place about 40–45 million years ago. The phylogeny and the closest relatives to ERV XA34 are discussed.

Introduction

Probably all eukaryotes carry endogenous retroviruses (ERVs) in their genomes. For some species the proportion of the genome consisting of ERVs is substantial, e.g. 2% for mouse and 1% for man (Gallo, 1995). So far, many human ERVs (Larsson et al., 1989; Lieb-Mösch et al., 1990) have been isolated and characterized, e.g. ERV-1 (Bonner et al., 1982), ERV-3 (O’Connell et al., 1984), ERV-9 (La Mantia et al., 1991), Humer 41 (Rabson et al., 1985; Repaske et al., 1985), RGH (Hirose et al., 1993) and RTVLH (Mager & Freeman, 1987). There is a great variability of copy number among different human ERVs. For instance, there are single or low copy ERVs like ERV-1 (Bonner et al., 1982) and ERV-3 (O’Connell et al., 1984) and there are others like Humer 41 (Rabson et al., 1985; Repaske et al., 1985) with 50 to 100 representatives and RTVLH (Goodchild et al., 1993; Mager & Freeman, 1987; Wilkinson et al., 1993) with more than 1000 copies per haploid genome.

ERVs enter the germ-line by means of infection and are then either lost or fixed in the population. For the large copy ERVs, the original integrations have probably been followed by a germ-line amplification by mechanisms other than infection (Wilkinson et al., 1990; Goodchild et al., 1993). ERVs represent new genetic material arriving at a specific chromosomal position at a specific time during evolution. The majority of ERVs that we can study in the human genome have been fixed. If the mutation (insertion of a retrovirus in the germ-line) does not affect the fitness, then fixation of the new genetic material will be a process of random drift. The time for fixation of a new endogenous retrovirus will then be 4N generations (Kimura, 1983), where N is the size of the population. Once fixed, it will take the same time for such a virus to be lost from the population by random drift. The localization and fixation at a specific chromosomal position of an ERV in a population or a species will be specific for that branch of population(s) or species. Hence, ERVs can be of great phylogenetic value.

In this paper we describe sequencing data and phylogenetic relationships of a novel family of ERVs. We have identified several related sequences of this new ERV, XA34, which was...
originally isolated from a human glioma cDNA library. By means of PCR we have isolated and analysed ≈ 940 bp pol fragments of three additional ERVs from human DNA that are closely related to XA34. A 4729 bp genomic clone representing an additional ERV of the XA34 ERV family with a 3' flanking region was also isolated and sequenced. The evolution of these ERVs is discussed.

Methods

■ PCR primers. The PCR primers used and referred to are:

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>390</td>
<td>5' TTATGAGATTTCTTCCAGGG 3'</td>
</tr>
<tr>
<td>490</td>
<td>5' ACCAAGTTAGCCGTGTTAAGT 3'</td>
</tr>
<tr>
<td>522</td>
<td>5' CCTGGAAGCCGCTCAGTAT 3'</td>
</tr>
<tr>
<td>523</td>
<td>5' ACCAAGCTTGTAATTGGTGC 3'</td>
</tr>
<tr>
<td>560</td>
<td>5' CAAAAACCGCTAGGCCTAGA 3'</td>
</tr>
<tr>
<td>561</td>
<td>5' GCCCAAAAGGGGATCTAAGCA 3'</td>
</tr>
<tr>
<td>593</td>
<td>5' CTGATGTCATTCTCAGTGG 3'</td>
</tr>
<tr>
<td>693</td>
<td>5' CAAGAATTCGGAGACCCTCTCG 3'</td>
</tr>
<tr>
<td>2793</td>
<td>5' TGTGAGGTAGCTATGGTCG 3'</td>
</tr>
</tbody>
</table>

■ DNA probes. An env probe from ERV-9 (La Mantia et al., 1991) was made by PCR with the primer pair 560/561. The reaction mixture (100 μl) contained 11 pmol of each primer, 0.1 μg human male total DNA, 2.5 mM-MgCl2, 0.2 mM-dNTP, standard buffer (Perkin-Elmer Cetus) and 1 U of AmpliTaq (Perkin-Elmer Cetus). The PCR was run for 30 cycles with a thermal profile of 95 °C for 45 s, 50 °C for 45 s and 72 °C for 60 s. The PCR generated a product of 250 bp which was of the expected size. This was further isolated from an agarose gel with glassmilk (Bio101).

A 378 bp pol fragment from XA34 and 142–155 bp pol fragments from XA34, XA35, XA36, XA37 and XA38 were used as probes for the hybridizations (Fig. 1). The 378 bp XA34 pol probe was made from the 5’ end of a 2303 bp XA34 CDNA fragment after digestion with EcoRV and EcoRI. The 142–155 bp pol probes from XA34, XA35, XA36, XA37 and XA38 were made by PCR from cloned material using primer pair 593/2793. The reaction mixtures (50 μl) contained 0.1 μg primer, 10 ng template. 2.5 mM-MgCl2, 0.2 mM-dNTP, standard buffer (Perkin-Elmer Cetus) and 1 U of AmpliTaq (Perkin-Elmer Cetus). The PCR products were separated and isolated from an agarose gel and the fragments were cloned in pTTBlue. The three clones were identical and after sequencing this human ERV was named XA38. The reason why we didn’t amplify XA38 from the 4.5 kb slice is probably because the primers didn’t fit this virus well enough. Primer 593 has four mismatches out of 22 bases. On the other hand, we probably would have picked up XA33 from the library of the 4.5 kb genomic EcoRI gel slice if we had screened more clones.

■ Sequencing and computer analysis. Dideoxy sequencing was performed according to Sanger et al. (1977), using Sequenase (USB) [a-35S]dATP (Amersham). Electrophoresis was done on wedge-shaped sequencing gels (0.2–0.4 mm thick). The software package (version 7.0) from the University of Wisconsin Genetics Computer Group (Devereux et al., 1984) was used for the DNA sequence analysis. The programs used within this package were Bestfit, FastA, TFastA, Lineup, Pretty, Blast, Pileup, Distances and Growtree. The trees were constructed using Pileup within this package were Bestfit, FastA, TFastA, Lineup, Pretty, Blast, Pileup, Distances and Growtree. The trees were constructed using Pileup to align the proteins. The phylogeny were then constructed using Distances with Kimura’s parameters and Growtree with either the neighbour-joining algorithm or the UPGMA algorithm. The EMBL or GenBank nucleotide databases were used for homology searches using FastA, TFastA or Blast. We use continuously updated versions of these databases.

■ Southern blot hybridization with total DNA. Total human DNA was prepared from blood lymphocytes. The chimpanzee (Pan troglodytes) female DNA and testis tissue were kind gifts from Ulfr Arnason at the Department of Evolutionary Molecular Systematics at the University of Lund. The orangutan (Pongo pygmaeus) material (blood lymphocytes from Anna and Dennis) was a kind gift from Jens Lilleor at the Aalborg Zoo in Denmark. The squirrel monkey (Saimiri sciureus) tissue (testis and kidney) was a kind gift from Thomas Brodin at the Department of Evolutionary Molecular Systematics at the University of Lund.
Pharmacia in Lund. Rat DNA was prepared from the liver of a male *Rattus norvegicus*, strain Wistar Furth. DNA was isolated by standard methods. Total DNA was digested with either *PstI* (Boehringer Mannheim) or *EcoRI* (Boehringer Mannheim); 15 µg of the digested DNA was run on a 0.8% agarose gel at 1.3 V/cm. After electrophoresis the gel was treated according to standard techniques (Ausubel et al., 1987). The DNA was transferred from the gel to a Biodyne B membrane, baked at 80 °C for 2 h and hybridized at 65 °C in Rapid Hybridization solution (Amersham) overnight. The concentrations of the labelled probes in the hybridization solution were 5 x 10^6 d.p.m./ml. The filters were washed with two changes of 2 x SSC--0.5% SDS for 1.5 h each at room temperature and then with 1 x SSC--0.5% SDS for 1.5 h at 65 °C. Autoradiography of the filters was typically done overnight. Before reprobing, the filters were stripped according to the manufacturer’s protocol and the blot was checked by autoradiography overnight at − 70 °C.

### Retroviruses

The accession numbers for the sequences used are as follows: baboon endogenous retrovirus (BaEV), M16550; ERV-3, M12140; ERV-9, X57147; feline leukaemia virus (FeLV), M18247; HSRIRT, M64936; HSRTVE, D10450; HUMER41, M10976; RDI1143, X51930; RESIV2DC, M16605; RESIVMPC, M12349; RGH1, D10083; RGH2, D11078; RTVHL2, M18048; SIVENVLTR, L38695; MLV, J02255, J02256, J02257; XA34, U29659; XA35, U37054; XA36, U37067; XA37, U29658; XA38, U37066.

### Results

Low stringency screening of the λ gt11 cDNA library from human glioma with the ERV-9 env probe yielded 50 positive clones. Seven of these clones were isolated and subcloned into the EcoRI site of M13mp18. Complete sequencing of four cDNA clones and partial sequencing of the three others showed that all were identical and came from the same ERV. The different clones varied in length from about 900 bp to 2303 bp but all had the 3' end with the poly(A) tail in common.

This ERV was called XA34 after the longest cDNA clone with a length of 2303 bp. Of the other 43 cDNA clones that scored positive after the low stringency probing, none were of ERV-9 origin and none could be identified as being of retroviral origin.

XA34 starts from the C-terminal end of the reverse transcriptase of the pol region followed by a partially truncated integrase and a truncated env region respectively that ends with a 3' Alu-LTR (Fig. 1). The somewhat truncated integrase ends with a stop codon at position 1817 followed almost immediately (position 1819) by an incomplete transmembrane (TM) protein. The integrase of XA34 is about 80 amino acids shorter than the corresponding protein of XA38, FeLV or BaEV. The env region, which completely lacks the surface (SU) protein, is also truncated at the C-terminal end of the TM region. The transmembrane protein region has strong homology with the ERV-9 TM protein. So, it is rather likely that recombination of XA34 and an ERV-9-like retrovirus has taken place at about position 1817 resulting in XA34. Further, a second event has probably occurred where the C-terminal end has recombined with an Alu repeat. Interestingly, the truncated XA34 env region contains the conserved region that has been...
implicated in immune suppression (Cianciolo et al., 1985). However, it lacks the SU region and the C-terminal part of p15E including the TM region. The presence of several frame shift and stop codon mutations in the coding regions makes it unlikely that the pol-env transcript would result in either Pol or Env peptides or proteins.

We performed Southern blot analysis of EcoRI-digested human DNA using a 378 bp XA34 pol probe (data not shown). This Southern blot revealed that XA34 is a multi-copy ERV. After high stringency washes the blot showed strongest hybridization with mainly three fragments. These three EcoRI fragments were about 4.5 kb, 7 kb and 20 kb in length. From a preparative gel three gel slices containing the above EcoRI-digested DNA were excised. PCR amplification of DNA isolated from the three different gel slices using primers 593/693 resulted in the amplification of 940 bp fragments in each case. The primer pair 593/693 was chosen from XA34 to amplify a reasonably large part of pol. Primers were also chosen from somewhat conserved regions after comparison of XA34 with other ERVs. The PCR fragments were cloned and sequenced. Besides XA34, the sequencing identified three additional retroviruses closely related to XA34 that were named XA35, XA36 and XA37 (Figs 1 and 2).

In an attempt to isolate the env region from a member of the XA34 family we made a small library from a 4.5 kb EcoRI gel slice in λ gt10. The small library was screened with the 378 bp XA34 pol probe and one positive clone of 4729 bp was further analysed. This ERV was, surprisingly, not identical to the clone XA35 that was identified after PCR of the 4.5 kb gel slice. This fifth ERV was designated XA38. The XA38 clone spanned from pol towards the 3' flanking side. The pol region spans from position 1 to position 1817 demonstrating an XA38 complete pol over this region.

The env region of XA34 is closely related to ERV-9 and shows almost no homology with the env region of the RGHs. This disparity, along with the missing splice acceptor site and gp70 region (see Fig. 3), clearly demonstrate that the env region of XA34 is probably the original env, although this region of the retrovirus seems to have mutated more than the pol region and the start of the env region is difficult to recognize. The env region is also slightly truncated at the C-terminal end. Shortly after the beginning of
the TM region the sequence has an Alu repeat starting at about position 3050 (Fig. 4). No obvious 3' LTR can be identified but three Alu repeats downstream from the region of a New

Computer analysis of the 940 bp

short probes were used for Southern blot analysis. Southern blot hybridization of the 142–155 bp probes from XA34–XA38 to PsI-digested male DNAs from man, chimpanzee, orangutan, squirrel monkey and rat are shown in Fig. 7.

The hybridization pattern shown in Fig. 7 is well conserved among the primates. The chimpanzee is most similar to the human but the orangutan also shows striking homology. After the rather high stringency washes (1 x SSC at 65 °C for 90 min) hybridization to DNA from squirrel monkey, representing a New World monkey, gives conspicuous bands. The positions of the bands in the hybridization to the squirrel monkey DNA are different from that for the chimpanzee and the orangutan. Considering the stringency of the washes and the conspicuous bands, we interpret the bands in the squirrel monkey to represent viruses common to those in the other primates and not of separate origin. The different positions of the bands in the squirrel monkey probably reflect mutations in the restriction enzyme site(s) resulting in alternative fragment lengths. If this is so, the introduction of the XA34 endogenous retroviruses in the primates took place between 35 to 45 million years ago, which is the estimated time of divergence between the Old and New World monkeys.
After hybridization with the specific 142–155 bp pol probes of XA34–XA38 (Fig. 7) bands that are not only cognate but also homologous can in some cases be resolved. It is obvious that, in the primate genomes, there are many more copies of endogenous retroviruses related to XA34 than just the five cloned members. In the hybridization with XA35 (Fig. 7) weaker double bands of about 3 kb in size are seen in both the human and the chimpanzee male DNAs, whereas only the shorter of these bands can be seen in the female DNAs. We propose that this band in the male represents an ERV on the Y chromosome. It is rather unlikely that this extra band should represent an RFLP that was recognized both in the human and the chimpanzee male DNA of different racial origin (data not shown).

From the hybridization with the XA34 probe (Fig. 8) weaker double bands of about 3 kb in size are seen in both the human and the chimpanzee male DNA, whereas only the shorter of these bands can be seen in the female DNAs. We propose that this band in the male represents an ERV on the Y chromosome. It is rather unlikely that this extra band should represent an RFLP that was recognized both in the human and chimpanzee male DNA. Moreover, we see this extra band in other human male DNA of different racial origin (data not shown). However, RFLP can be seen and one example of this is the
Fig. 4. Sequence of the XA38 4729 bp genomic EcoRI fragment. The pol region from position 1–1817 is underlined ending with a stop codon (marked with an asterisk). The env region follows pol and a region of the env at the end of the SU region and the beginning of the TM region is translated. Three Alu repeats that follow the truncated env are underlined.
Fig. 5. Phylogenetic relationships between Env proteins of XA38, RTVLH-RGH1, RTVLH-RGH2, ERV-3, HSRIRT, HUMER4-1, HSRTVE, BaEV, RD1143, RESIVMPC, SIVENVLTR and RESIV2DC given as a tree. The aligned protein of XA38 is the translation from Fig. 4 and the corresponding proteins from the other retroviruses. The programs used for aligning the sequences were Pileup and Lineup. The phylogeny was reconstructed using the programs Distances and Growtree. The Kimura protein distance correction algorithm was used together with the UPGMA algorithm.

Fig. 6. Phylogenetic relationships between Pol proteins and deduced Pol proteins of XA34, XA35, XA36, XA37, XA38, RTVLH2, RTVLH-RGH1, RTVLH-RGH2, ERV-9, HUMER4-1, MLV, FeLV and BaEV given as a tree. The aligned sequences of the XA34 family can be seen in Fig. 2. The Pol proteins for the other retroviruses are the corresponding sequences. The programs used for aligning the sequences were Pileup and Lineup from University of Wisconsin Genetics Computer Group (Devereux et al., 1984). The phylogeny was reconstructed using the programs Distances and Growtree. The Kimura protein distance correction algorithm was used together with the UPGMA algorithm.

---

Discussion

We describe the isolation and characterization of the human ERVs XA34–XA38. The evolutionary relationships of these retroviruses (Fig. 5) imply that they should be classified as a new ERV family.

For XA34, DNA sequence comparisons suggest that this virus has recombined with an ERV-9-like retrovirus at about position 1817 and during the process of recombination lost the C-terminal end of pol and most of the ERV-9 env region (Fig. 3). In order to determine the correct structure of the env region for the XA34 family of endogenous retroviruses we cloned and sequenced the 3’ flanking region of the related XA38. This virus had a complete pol region when compared with BaEV and

---

\[ \approx 5 \text{ kb band and the } \approx 4.4 \text{ kb band seen in the human male and female DNA respectively. The female used in this blot is the only female where this shorter band of 4.4 kb has been detected. The human and chimpanzee male DNA 3 kb band cannot be detected in the orangutan. However, since the homology between this virus present on the Y chromosome and XA34 is more limited we cannot decide whether this virus is present or not in the orangutan lineage although an extra band of about 2.3 kb can be detected in the male orangutan. We have previously reported that some endogenous retroviruses are over-represented on the Y chromosome (Kjellman et al., 1995). The hybridization to the female chimpanzee DNA is somewhat fuzzy because this DNA was of lower molecular mass and it was not possible to obtain good signals from longer fragments without loading about double the amount of DNA. This overloading of the female chimpanzee DNA led to bands that are somewhat out of position and, hence, give stronger signals with lower molecular mass DNA. We kept this lane in Fig. 8 as we think that it is still informative.\]
FeLV. The XA38 env region is somewhat difficult to identify but there is a region from position 2827 to 3000 corresponding to the end of the SU and the beginning of the TM region that can be tracked. This region shows the strongest homology with the RGHs (Fig. 6) which makes it plausible that this is a true env region.

This env region is not complete and starting at position 3050 we can identify an Alu repeat. This Alu repeat is followed by two additional ones and this makes it likely that the original XA38 has lost the end of the env and the 3' LTR completely.

The Southern blot analyses in Fig. 7 and the sequence analyses of pol fragments (Fig. 6) from five members of the XA34 family show that the elements within the family are closely related to each other. From the Southern blot analyses in Figs 7 and 8 it can be seen that the XA34 ERV family is present in all primate lineages tested. Although lemurs and other more primitive apes were not studied, 12 related retroviruses were detected in the squirrel monkey, a New World ape. These 12 bands in the squirrel monkey are difficult to identify in Fig. 7, where no more than seven bands can be distinguished, but after prolonged exposure 12 bands are detectable. Hence, XA34 retroviruses probably entered the primate genome about 40–45 million years ago. The separation of the great apes (humans, chimpanzees, gorillas and orangutans) from the monkeys took place about 35 million years ago and was preceded by the separation of New World and Old World monkeys that took place about 40–45 million years ago. It is difficult to decide whether XA34 was introduced in the genome as one single insertion and later amplified or if multiple infections of the germ-line have taken place.

It is likely that at least one copy of an XA34 related endogenous retrovirus is located on the Y chromosome.
Expression of sequences on the Y chromosome is very limited, especially expression of sequences within or in the vicinity of the larger heterochromatic part of this chromosome. We argue that an over-representation of endogenous retroviruses (see Kjellman et al., 1995) on the Y chromosome is mediated by three main factors: (i) absence of recombination of the Y chromosome that makes it more difficult for sequences to be lost; (ii) integration of retroviruses in or near heterochromatic regions is less harmful to the organism and (iii) the time for fixation on the Y chromosome is four times faster within a population that has an equal sex ratio. The normal time for fixation of a neutral mutation is $4N$ generations but for fixation on the Y chromosome this figure will be $2N$ generations, and if $N$ is the population size of males alone, this figure will be about half the normal size.

XA34 was first isolated as a cDNA from a glioma cDNA library. It is intriguing as to why this virus is expressed in this malignant tissue. A study has been initiated to analyse the expression of endogenous retroviruses of the XA34 family in a range of malignant and normal tissues (C. Kjellman and others, unpublished). What we can say so far is that the XA34 virus is abundantly expressed in normal human placenta. Several endogenous retroviruses have been reported previously to be expressed in the placenta (Kato et al., 1987; Lyden et al., 1994; Rabson et al., 1985). It is noteworthy that if an endogenous retrovirus that is present only on the Y chromosome is expressed and translated to a functional peptide, such a peptide would represent an HY-antigen (Wiberg, 1987) by definition. It is also well known that females, after giving birth to several male offspring, have antibodies
that react with male cells. Some of these HY-antigens could simply be of retroviral origin, having nothing to do with sex differentiation.

We would like to thank Ms Ingar Nilsson for skilful technical assistance. This work was supported by the Swedish Cancer Foundation, the Swedish Medical Research Council, the Nilsson-Ehle Foundation, the Blücher Foundation, the John and Augusta Persson's Foundation and the Medical Faculty of the University of Lund.

References


Received 8 January 1996; Accepted 29 March 1996