An unusual retrovirus-like sequence identified in human DNA

Laura S. Levy, Patricia A. Lobelle-Rich, John H. Elder, Susan Payne and Ronald C. Montelaro

Department of Microbiology and Immunology, Tulane University School of Medicine, 1430 Tulane Avenue, New Orleans, Louisiana 70112, Scripps Clinic and Research Foundation, La Jolla, California 92037 and Department of Biochemistry, Louisiana State University, Baton Rouge, Louisiana 70803, U.S.A.

The human genome contains many different types of endogenous proviruses and retrovirus-like elements. An unusual element of this kind has been isolated from human DNA on the basis of its relatedness to the integrase-coding domain of the pol gene of feline leukaemia virus (FeLV). The element, termed Hs5, is related to FeLV only over a short region of 81 nucleotides predicted to encode the carboxyl terminus of the FeLV integrase protein, p46pol. The region of relatedness between Hs5 and FeLV identifies a short conserved amino acid stretch which is shared among distantly related retroviruses. The conservation of this sequence, its position, and predicted secondary structure suggest that it may represent a conserved substrate binding site or active site of the integrase enzyme. Nucleotide sequence analysis of Hs5 reveals that it is not an intact retrovirus, but contains only the 3' terminus of pol and a defective env gene without apparent long terminal repeat; Hs5 is unusual among human endogenous retrovirus-like elements in this respect.

A unique feature which distinguishes the Retroviridae among animal viruses is that the DNA of normal, uninfected animal cells contains gene sequences which are closely related to exogenous, infectious retroviral species. Indeed, the human genome contains many different types of endogenous proviruses and retrovirus-like elements. Such sequences have been identified by hybridization at low stringency with probes representing murine (Callahan et al., 1982; Repaske et al., 1983; Steele et al., 1984; Deen & Sweet, 1986; Ono, 1986) or primate (Noda et al., 1982; O'Connell et al., 1984; Leib-Mosch et al., 1986) proviral DNAs, or with oligonucleotide probes complementary to retroviral primer binding sites (Harada et al., 1987; Kröger & Horak, 1987). Other elements of this type have been identified fortuitously by their proximity to unrelated gene sequences (Mager & Henthorn, 1984; Mager & Freeman, 1987), or by polymerase chain reaction using primers homologous to retroviral reverse transcriptase (Shih et al., 1989). Some of these sequences are transcribed in a tissue-specific manner (Rabson et al., 1985; Kato et al., 1987), although no human tissue has been isolated which is productively infected with an endogenous retrovirus. The function of endogenous retroviruses and retrovirus-like elements in human DNA is not known, but several possibilities have been suggested (Temin, 1985). Endogenous retroviruses may have played a role in evolution of the human genome, e.g. by serving as a source of genetic variation, by increasing the effective target size for possible lethal mutations, or by acting as insertional mutagens. In their present form, they may serve as a reservoir for recombination and may thereby contribute to the origin of pathogenic retroviruses. We report here the identification of an unusual retrovirus-like element isolated from human DNA by virtue of its relatedness to feline leukaemia virus (FeLV), a horizontally transmissible retrovirus of the domestic cat.

Southern blot analysis of normal human DNA was performed using as probe a 3-6 kb KpnI fragment of FeLV proviral DNA (Fig. 1a). The probe contains the 3' end of the pol gene (1-2 kb), the entire env gene (2-0 kb), and the U3 region of the FeLV long terminal repeat (LTR) (0-4 kb; Mullins et al., 1981; Elder & Mullins, 1983). This fragment was used as a probe because it contains two regions of interest. First, it contains the region of pol which encodes the integrase function (Tanese et al., 1986). The integrase domain is known to be highly conserved among retroviruses (McClure et al., 1988), but has not been used in this manner as a hybridization probe to identify related sequences in...
human DNA. Second, certain human tumours produce an inhibitor of monocyte chemotaxis which is recognized by monoclonal antibodies to p15E, an envelope protein of FeLV (Cianciolo et al., 1984b). Thus, a probe containing the FeLV env gene might identify the gene which encodes the chemotaxis inhibitor. Hybridization was performed under conditions of moderate stringency at 37°C in a buffer containing 50% formamide, 1 M NaCl, 0.1% sodium lauryl sarcosine, 0.01 M EDTA, 0.2 mg of denatured calf thymus DNA per ml, and 5 × Denhardt’s solution, buffered to pH 6.8 with 0.05 M PIPES. Non-specifically bound probe was removed by washing twice for 30 min each at 37°C in 2 × SSC (1 × SSC is 0.15 M NaCl and 0.015 M sodium citrate) with 0.1% sodium lauryl sarcosine, and twice for 30 min each at 45°C in 0.1 × SSC with 0.1% sodium lauryl sarcosine. When the 3.6 kb KpnI fragment of FeLV was used in Southern blot analysis under these conditions, multiple fragments were identified in human DNA (data not shown).

A recombinant library of genomic DNA from human sperm was constructed in the bacteriophage cloning vector λEMBL3 (Frischauf et al., 1983) as previously described (Levy et al., 1984). The library (1 × 10^6 recombinants) was screened by in situ plaque hybridization using the FeLV probe. Nineteen recombinants were selected from the library, none of which was recognized by a 2.5 kb PstI–KpnI fragment of the probe (Fig. 1a) which represents FeLV p15E env and U3. All of the selected recombinants were recognized by that portion of the probe containing the 3′ pol sequences. Restriction enzyme site mapping demonstrated some of the recombinants to belong to a previously identified family of endogenous retroviral sequences, the 4-1 family (Repaske et al., 1985; data not shown). The restriction enzyme site map of one recombinant, Hs5, did not resemble that of 4-1 (Fig. 1b), and that recombinant was selected for further study.

The entire relatedness to FeLV contained in Hs5 was shown to reside in a 0.4 kb BamHI fragment, designated Hs5.4. The nucleotide sequence of Hs5.4 was determined using dideoxynucleotide chain termination in

M13-based vectors. Computer-assisted analysis of the sequence data was performed using Microgenie (Beckman Instruments) or Pustell Programs (International Biotechnologies). The sequence of Hs5.4 was compared to that of a 679 nucleotide HindIII–PstI fragment of FeLV (indicated in Fig. 1a) which contains the 3′ end of FeLV pol and thus the relatedness to Hs5.4. Comparison (Fig. 2) demonstrates that Hs5.4, FeLV and 4-1 are related to each other over a region of only 81 nucleotides at the 3′ end of Hs5.4. Within this region, all three sequences are identical at 61 of 81 positions (75%), and Hs5.4 is identical to FeLV at 68 of 81 positions (84%).

Fig. 1. (a) Partial restriction enzyme site map of proviral DNA of the Gardner-Arnstein-B strain of FeLV (Mullins et al., 1981). Indicated is a 3.6 kb KpnI fragment used as hybridization probe. Also indicated is a 679-bp HindIII–PstI fragment whose nucleotide sequence was determined. (b) Restriction enzyme site map of Hs5 DNA. Indicated is a 0.4 kb BamHI fragment termed Hs5.4. The region of Hs5 whose nucleotide sequence was determined is underlined; the predicted direction of transcription is indicated by an arrow. Restriction enzymes: B, BamHI; H, HindIII; K, KpnI; P, PstI; RI, EcoRI; Sa, SalI; X, XbaI. The SalI recognition sites at both ends of Hs5 DNA are contributed by the cloning vector.
Short communication

Fig. 3. Predicted translation of the carboxyl terminus of pol as encoded by Hs5, 4-1, FeLV, marine leukaemia virus, simian sarcoma virus, reticuloendotheliosis virus A, human T cell leukaemia viruses, bovine leukaemia virus, human immunodeficiency virus, equine infectious anaemia virus, visna virus, human spumaretrovirus and simian foamy virus type I (Repaske et al., 1985; Sonigo et al., 1985; Weiss et al., 1985; Rushlow et al., 1986; Fliegel et al., 1987; Mergia et al., 1990). The box denotes a conserved six amino acid sequence. The SI designation following the predicted translation of HSRV indicates that the peptide is found not at the pol terminus, but within an open reading frame in the intergenic region between pol and env (Fliegel et al., 1987).

Over the remaining 287 nucleotides, the three sequences are not significantly related (Fig. 2). A search of the GenBank DataBase does not demonstrate relatedness of Hs5.4 to any other sequence.

The 81 nucleotide region within which FeLV, Hs5.4, and 4-1 are related is located in the domain of FeLV pol which is predicted to encode the carboxyl terminus of the integrase protein, p46pol (Johnson et al., 1986; Tanese et al., 1986). A six amino acid stretch in the predicted translation of this region is identical among the three, and is strongly conserved among other retroviruses which are only distantly related. The six residue sequence itself is conserved, as is its position in the terminal 70 residues of the integrase protein (Fig. 3). The consensus sequence, PRWKGP, is composed of two proline residues flanking four alternating charged and neutral residues; thus, it is likely to represent a bend in the secondary structure exposed at the protein surface (Janin, 1979; Schultz & Schirmer, 1979). In fact, the relative hydrophilicity of the region is evident even in sequences in which the consensus bears substitutions. Hydrophathy analysis of a portion of the pol protein including the region of interest illustrates that it represents part of a hydrophilic domain immediately flanked on the carboxyl side by a hydrophobic stretch (Fig. 4). The conservation among distantly related retroviruses of this sequence, its position, and predicted secondary structure suggest that the domain may represent a conserved substrate binding site or active site of the integrase. Mutants constructed in murine leukaemia virus pol have demonstrated a domain including the six amino acid sequence to be essential for integrase function, but the six amino acid sequence itself has not been targeted for mutation in these studies (Goff & Lobel, 1987; Donehower, 1988).

The sequence of 4582 nucleotides of Hs5 was determined (Fig. 5) and of the first 219 nucleotides, 216 are adenosine or thymidine residues (98.6% A + T). The following 1909 residues (positions 220 to 2128) were demonstrated by computer-assisted search of the GenBank DataBase not to be significantly related to any known retrovirus or other gene sequence. Predicted six amino acid sequence to be essential for integrase function, but the six amino acid sequence itself has not been targeted for mutation in these studies (Goff & Lobel, 1987; Donehower, 1988).

The sequence of 4582 nucleotides of Hs5 was determined (Fig. 5) and of the first 219 nucleotides, 216 are adenosine or thymidine residues (98.6% A + T). The following 1909 residues (positions 220 to 2128) were demonstrated by computer-assisted search of the GenBank DataBase not to be significantly related to any known retrovirus or other gene sequence. Predicted
Fig. 5. Sequence of 4582 nucleotides of Hs5. Predicted translation of Hs5 is indicated, aligned with the predicted translation of 4-1 across a homologous region (Repaske et al., 1985). Frameshifts, denoted with double underline, and gaps have been inserted to optimize the alignment. Mismatches between the two sequences are indicated (*). Also indicated are predicted glycosylation sites of 4-1 (Repaske et al., 1985), and the 26 amino acid (underline) and 36 amino acid (dotted underline) regions of conservation with pl5E as described in the text.
translation of the sequence demonstrates multiple termination codons in all reading frames. The longest potential open reading frame encoded by this portion of the sequence is 554 nucleotides long (positions 854 to 1408). The following 2454 residues (positions 2129 to 4582) are distinctly retrovirus-like, and demonstrate strong relatedness to the 3' terminus of pol and the env gene of 4-1. Computer-assisted alignment demonstrates 77% identity between the sequences over a region of Hs5 including positions 2129 to 4442. At optimum alignment, the predicted translations of Hs5 and 4-1 are identical at 545 of 762 amino acid residues (71.5%). Hallmarks of the 4-1 sequence conserved in Hs5 (indicated in Fig. 5) include a 22 base pair spacer between the predicted end of pol and start of env, positions of potential glycosylation sites, a 26 amino acid stretch in p15E conserved among mammalian retroviruses (Cianciolo et al., 1984a), and a hydrophobic 36 amino acid sequence near the carboxy terminus of p15E, followed by lysine and a hydrophilic region. The latter is predicted to represent the transmembrane domain of p15E (Lenz et al., 1982). Sequence relatedness between Hs5 and 4-1 ends abruptly at position 4442. The remaining 140 nucleotides of Hs5 (positions 4442 to 4582) are not related to the 4-1 LTR or to any LTR sequence reported to the GenBank Database. Thus, Hs5 appears to contain only the 3' terminus of pol and a defective env gene without LTR sequences, flanked on both sides by cellular DNA.

A 1.5 kb BamHI restriction fragment of Hs5 isolated from the region of relatedness between Hs5 and 4-1 (positions 3071 to 4570; Fig. 5) was subcloned and used as a probe in Southern blot analysis of cloned 4-1 DNA (a gift from Dr Arifa Khan). This analysis demonstrated that Hs5 may be distinguished from 4-1 only under conditions of high hybridization stringency, as might be predicted from their sequence relatedness across the region contained in the probe. Specifically, high stringency hybridization was performed in a buffer containing 2 × SSC, 10 × Denhardt's solution, 0.1% SDS and 0.2 mg denatured calf thymus DNA per ml at 68 °C for 16 h. Following hybridization, non-specifically bound probe was removed by washing six times for 5 min each at 68 °C in 2 × SSC with 0.1% SDS and three times for 15 min each in 0.1 × SSC with 0.1% SDS (data not shown). Under the same conditions of hybridization, the Hs5 probe failed to recognize 4-1 transcripts in RNA from human kidney (A. Khan, personal communication).

Hs5 and 4-1 exhibit sequence relatedness across the env region, but Hs5 is distinct in structure from 4-1 and other reported members of the 4-1 family. Truncated members of the 4-1 family have been described which contain terminal repeat, gag and pol sequences (Repaske et al., 1983; Steele et al., 1984), but none has been reported which, like Hs5, contains isolated env. Although its possible function is not known, a sequence of this type might represent a reservoir for recombination and thereby contribute to the origin of new pathogenic retroviruses. Another possible role is suggested by the similarity of this structure to the murine Fv-4 element, since both contain cellular sequences flanking a truncated, 3' proviral element. The Fv-4' phenotype is associated with resistance to infection by ecotropic murine leukaemia viruses, a property proposed to result from interference based on expression of env sequences at the Fv-4 locus (Ikeda et al., 1985). Analysis of other members of the Hs5 family and of their transcriptional pattern could reveal whether Hs5 expression serves a similar function.

The authors acknowledge Dr Arifa Khan for communicating unpublished information, for providing cloned 4-1 DNA, and for helpful discussions. L.S.L. acknowledges support of the Louisiana Board of Regents (Contract no. 84-TUM/010-B2), the American Cancer Society (IN-133G) and the Ladies Leukemia League of New Orleans.

References


(Received 9 October 1989; Accepted 22 February 1990)