Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 4709–4713, August 1982
Genetics

Cloned endogenous retroviral sequences from human DNA

(recombinant DNA/retrovirus evolution/DNA sequence determination)

T. I. BONNER*, C. O'CONNELL†, AND M. COHEN†

*Laboratory of Viral Carcinogenesis, National Cancer Institute, Frederick, Maryland 21701; and †Biological Carcinogenesis Program, NCI-Frederick Cancer Research Facility, Frederick, Maryland 21701

Communicated by Norman Davidson, May 4, 1982

ABSTRACT **We have screened a human DNA library using as probe a chimpanzee sequence that contains homology to the polymerase gene of the endogenous baboon virus. One set of overlapping clones spans about 20 kilobases and contains regions of DNA sequence homology to the *gag* p30, *gag* p15, and polymerase genes of Moloney murine leukemia virus. Furthermore, the spacings are the same as in Moloney virus between these sequences and a 480-nucleotide region that has the structural characteristics of a 3′ copy of the long terminal repeat sequence. Hybridization of the cloned DNA to restriction digests of human DNA indicates that the human genome contains only two copies closely related to the sequence and ≈10 less closely related copies. This retroviral sequence appears to have been in its present chromosomal location prior to the divergence of man and chimpanzee because the human and chimpanzee clones have 3–4 kilobases of identical 3′ flanking sequence.**

Retroviruses have been shown to be capable of inducing tumors by two basic mechanisms. The first involves the incorporation of a cellular transforming gene, or oncogene, into a retrovirus so that it is expressed whenever the virus infects cells (1). The number of oncogenes in a mammalian genome is uncertain but is clearly more than one (2). The second mechanism, demonstrated by the induction of bursal lymphomas in chickens by avian leukosis virus, involves the ability of the retroviruses to integrate a DNA copy of the viral genome into the chromosomal DNA of the host. In the rare case in which the viral DNA is integrated adjacent to an oncogene, the promoter sequences present in the termini of the integrated virus can activate the oncogene (3).

As a result of the ability of the retroviruses to integrate into host DNA and, thus, become part of the host genome, many, if not most, mammals have accumulated tens or hundreds of integrated retroviral sequences. These endogenous viral sequences thus provide a reservoir of viral genes which might be activated by developmental or environmental factors including mutagens. These newly activated viruses then might induce tumors either by the mechanisms described above or through other mechanisms, such as inactivation of a critical gene by integration within the gene (4). Evidence that the activation of endogenous viral genes may be significant is provided by the observation that formation of a recombinant virus derived from two endogenous viruses is correlated with spontaneous leukemias in the AKR mouse (5), a strain with a high incidence of leukemia.

The critical questions concerning retroviruses are whether they are involved in human carcinogenesis and, if so, to what degree. Attempts to answer these questions have been hampered by a dearth of human retroviruses. There is currently one human retrovirus (6). This virus is not an endogenous human retrovirus because related sequences are not found in normal human DNA (7). There is also a potential endogenous human retroviral sequence that was cloned from human DNA (8). To date this clone has been shown only to contain sequences homologous to the retroviral polymerase gene. In this report we describe the cloning of an unrelated endogenous human retroviral sequence that has homology not only to a retroviral polymerase gene but also to the *gag* genes p15 and p30. This sequence also has an apparent long terminal repeat (LTR).

MATERIALS AND METHODS

Nucleic Acid Hybridization. ^{32}P-Labeled probes were prepared by nick-translation of plasmid DNAs containing the appropriate viral sequences at specific activities of 2×10^7 to 2×10^8 cpm per μg. Hybridization to nitrocellulose filters to which DNA was fixed, either through plaque lift procedures (9) or Southern blotting (10), was as described (11) except that both hybridization and wash were in 0.45 M NaCl/0.045 M Na citrate, pH 7, at 60°C for low stringency and washes were in 0.03 M NaCl/0.003 M Na citrate, pH 7, at 60°C for higher stringency.

Computer Alignment of Sequences. The sequence alignment of distantly related DNA sequences was determined with the ALIGN program (12) by using a break penalty of 15 and a scoring matrix that gives 10 points for a nucleotide match, 3 points for a C-T or A-G mismatch, and 1 point for any other mismatch. The scoring for C-T and A-G mismatches was used to enhance the detection of distant homologies because the analysis of sequence data (13, 14, 15) indicates that transitions, that is C-T or A-G changes, are 2 to 3 times more common than transversions. To judge the significance of a particular alignment, 300 random reshufflings of the sequence were performed. The score of the particular alignment relative to the mean of the random scores is expressed in units of the standard deviation of the random scores. Assuming that the scores of all random sequences would be normally distributed about this mean with the same standard deviation, we can convert this alignment score to a probability of chance occurrence by using standard probability tables.

RESULTS

The Cloning of Human Sequences. A human library (16) was screened by using a nick-translated fragment (Fig. 1, map positions 12.6–13.35) of the chimpanzee clone CH2. This clone was obtained while cloning chimpanzee sequences related to the endogenous colobus retrovirus CPC-1 (17) but had little homology to CPC-1. However, this fragment has homology with the polymerase gene of the baboon endogenous virus (BaEV) (unpublished data). We obtained the clone HC-20 (Fig. 1) in

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: BaEV, baboon endogenous virus; kb, kilobase(s); LTR, long terminal repeat; M-MuLV, Moloney murine leukemia virus.

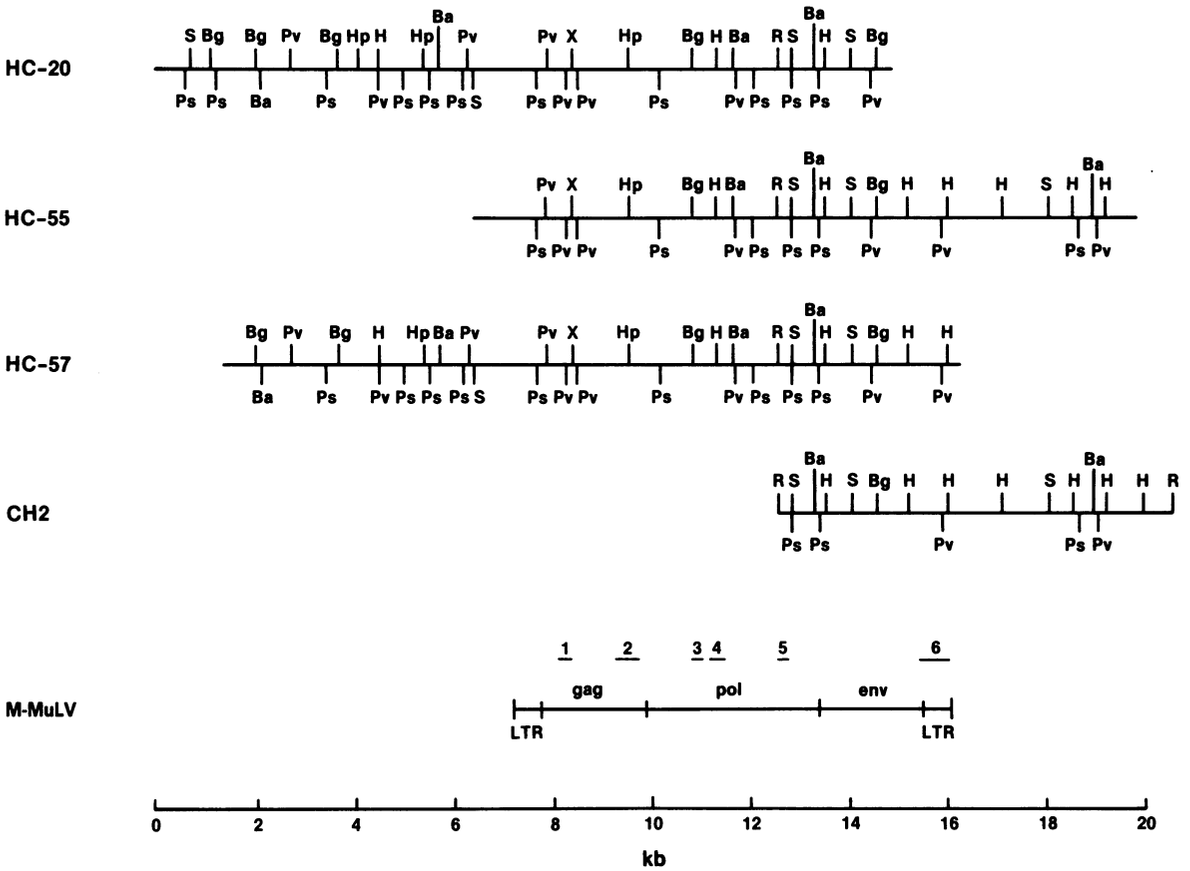

FIG. 1. The alignment of restriction maps of the human clones HC-20, HC-55, and HC-57 and the chimpanzee clone CH2. Ba, *Bam*HI; Bg, *Bgl* II; H, *Hin*dIII; Hp, *Hpa* I; Ps, *Pst* I; Pv, *Pvu* II; R, *Eco*RI; S, *Sac* I; X, *Xba* I. The *Eco*RI sites at the ends of the human clones are not shown because they are due to linkers used in constructing the library and, therefore, do not occur in human DNA. The genetic map of M-MuLV is shown at the bottom as it aligns with the restriction map by sequence determination of the numbered regions.

the initial screening. Upon rescreening the library with the 2.3-kilobase (kb) *Eco*RI fragment of HC-20 (Fig. 1, map positions 12.6–14.9) at high stringency, we obtained six more clones. Two of the new clones, HC-55 and HC-57, in combination with HC-20 represent 20 kb of a single genomic locus because they contain an overlapping set of identical restriction sites. Two of the other clones were identical to HC-57. Because four of the six new clones represented the same locus, the human library appears to contain few copies of sequence closely related to HC-20. Also shown in Fig. 1 is the restriction map of the chimpanzee clone CH2. Except for an additional *Pvu* II site in the human clones, the restriction maps of CH2 and the human clones are identical, indicating that they are closely related.

Identification of Retroviral Genes. To detect retroviral homologies that are too distant to be detected by low-stringency hybridization, we determined the sequence (18) of portions of the human clones. We have compared the resulting sequences to the sequence of Moloney leukemia virus (M-MuLV) (19) because this was the only complete retroviral sequence that was available. However, the available BaEV sequence proved to be only slightly more closely related.

Nearly 200 nucleotides of the CH2 and HC-20 sequences at site 5 of Fig. 1 and the corresponding BaEV sequence were determined. They are shown in Fig. 2 aligned with the M-MuLV virus sequence. As was indicated by their conserved restriction sites, the human and chimpanzee clones are closely related, having 98% nucleotide homology. The human and BaEV sequences are 62% homologous, whereas the human and M-MuLV sequences are 57% homologous. In contrast, the BaEV and M-MuLV sequences are 69% homologous. Thus, the human sequence is almost equally related to BaEV and M-MuLV viruses but less related to these two viruses than they are to each other. The 57% homology to M-MuLV is clearly significant because the ALIGN program gave a probability of 10^{-26} of a random sequence having this degree of alignment. Thus, the probability that such a sequence would occur in the 3×10^9 nucleotides of the human genome is $<10^{-16}$. Even stronger homology is evident within this sequence between nucleotides 5,013 and 5,085 of M-MuLV. In this region, the human sequence is 77% homologous to BaEV and 68% homologous to M-MuLV and is probably responsible for the hybridization of HC-20 and CH2 to the BaEV polymerase gene. An additional 450 nucleotides of polymerase homology was obtained at sites 3 and 4 (Table 1). Combined, these three sites indicate that conserved polymerase homologies extend over 2.0 kb.

We also have demonstrated the presence of *gag* gene sequences in the human clones through DNA sequence determinations at sites 1 and 2 of Fig. 1. The homology at site 1 to the M-MuLV p15 gene is sufficiently significant that the probability of its occurring by chance in the human genome is only 3×10^{-4} (Table 1; Fig. 3). The p30 homology at site 2 (Table 1; Fig. 4) is not sufficiently close to M-MuLV to be considered significant when compared to the 3×10^9 possible sequences in the human genome. However, strong evidence of the significance of the p30 homology can be obtained by considering the amino acid sequence encoded by the human DNA sequence. It is clear that the M-MuLV sequence maintains the correct reading frame throughout the p30 gene because the complete amino acid sequence of Rauscher murine leukemia

```
                    4958                          4975
Chimp     AATTCAGGAAGTTGGAGCAG  CGCCTTGTGAGAACCTACTT
Human                   ..      .A..................
BaEV            AACCCGT.GCAGC   .A..CG.---.GG.T..T.G
M-MuLV    ..C.-...GTCCGC.G...T  ..G.CC.---.C..TC.T.G

                    4995                          5015
Chimp     G----TGGACTTTACCAAAT  T---GCCTCTGGTCGGGGGC
Human     ...................   ...........-........
BaEV      .GAGG.A.....C..TG..G  .AAAA....ACTAT.CT..G
M-MuLV    .GAGA.C..T..C...G.GA  .AAA...CGGAT.GTAT...

                    5035                          5055
Chimp     TATCGGTACATGTTGGTGTT  TGTCTGCACCTTTTCAGGAT
Human     ...................   ....................
BaEV      ...AA....T.AC.A.....  ...AGAT.............
M-MuLV    ...AAA..TC.TC.A..T..  .A.AGAT........T..C.

                    5075                          5095
Chimp     GGGTAGAGGCTTTCCCCACC  CAAACAGAAAAGGAACAAGA
Human     ...................   ............C......
BaEV      .......A..C.........  .GGCA.....C..C...CAT
M-MuLV    ..A....A..C.....A...  A.G.A.....CC.CCA.GGT

                    5115                          5135
Chimp     GGTAACCCAGGTGTTGCTAA  GAGACATTATTCCCAGGTT
Human     ...................   ...............
BaEV      A...G..A..AA.A.C...G  A..
M-MuLV    C......A..AA.C.A...G  AG..G..CT.C........C
```

FIG. 2. Sequence comparisons at site 5 in the polymerase gene. The sequence of the human clone and chimpanzee clone are aligned with the corresponding sequences of M-MuLV and BaEV. The complete chimpanzee sequence is shown, and the other sequences are shown only where they differ from the chimpanzee sequence: ·····, the nucleotides are the same as in chimpanzee; ---, deleted nucleotides. The numbers identify positions in the M-MuLV genome (19).

virus (S. Oroszlan, personal communication) is nearly identical to the sequence predicted by the M-MuLV DNA sequence. Therefore, we have inferred the reading frame of the human DNA sequence from its alignment with the M-MuLV DNA sequence as shown in Fig. 4. The sequences are homologous at

Table 1. Alignment of human sequences with M-MuLV

Human site*	Gene	Nucleotide identities†	Alignment score/ probability‡	Spacing from previous site: Human clones§	Spacing from previous site: M-MuLV¶
1;5′ side of *Xba* I	*gag* p15	57% (73/127)	$7.5/10^{-13}$	0	0
2;5′ side of *Hpa* I	*gag* p30	51% (95/185)	$6.6/10^{-10}$	1.1	1.05
3;3′ side of *Bgl* II	*pol*	58% (72/124)	$9.9/10^{-22}$	2.35	2.45
4a;5′ side of *Hin*dIII	*pol*	57% (104/182)	$8.8/10^{-18}$	0.49	0.49
4b;3′ side of *Hin*dIII	*pol*	57% (84/147)	$9.8/10^{-22}$		
5;3′ side of *Eco*RI	*pol*	57% (103/180)	$10.7/10^{-26}$	1.26	1.25
6;presumptive transcription termination in 3′ LTR				3.35	3.39

* The site in the human sequence is identified by the site number from Fig. 1 and the restriction site which was labeled.

† Total nucleotides include all deleted nucleotides inferred from the sequence alignment.

‡ The alignment score and the probability of a random sequence giving this degree of alignment were derived as described in *Materials and Methods*.

§ The spacing between the corresponding restriction sites.

¶ The spacing between the M-MuLV sequences homologous to the human restriction sites. Because the sequence was not read through the human restriction site in most cases, the position is estimated with an error of about ±20 nucleotides.

```
                     762                           782
Human     AAAATGGCCTTCTTTTNATG  TCGGGTGGCCAGCTGAAGGA
M-MuLV    .G.......AA.C...A.C.  ....A.....GCGA..C..C

                     802                           822
Human     ACAATAGATAGGGAAGCAAT  TGGCCATGT---GTTCAGGG
M-MuLV    ..CT.TA.CC.A..CCTC..  CAC...G..TAA.A...A..

                     836                           856
Human     TAGTAACCGGAGTTGGAGGA  CAGCCTGAGCACCCAGATCA
M-MuLV    .CT.TT.---.CC..---.C  .C..A..GA........C..

                863
Human     GTTTCCA
M-MuLV    .G.C..C
```

FIG. 3. Alignment of human sequences at site 1 with the *gag* p15 gene of M-MuLV. DNA sequences of HC-20 and M-MuLV are presented as in Fig. 2.

25 of 51 amino acids between M-MuLV nucleotides 1,706–1,859. These amino acid sequences gave an alignment score (13) of 12, indicating a probability of $<10^{-32}$ that this homology would occur by chance.

Although the p15 *gag*, p30 *gag*, and polymerase homologies are independently significant when considered in the context of the whole human genome, they occur in a 5.5-kb segment of human DNA with the same spacings as in M-MuLV. Thus, once the significance of one sequence, for example, the polymerase sequence at site 5, has been judged relative to the whole human genome, the significance of the remaining sequences should be judged relative to a much smaller number of nucleotides. The probabilities then become so small that we must conclude that these homologies reflect the presence in humans of retroviral *gag* genes in association with a retroviral polymerase gene.

A Possible LTR Sequence. Another feature characteristic of integrated retroviruses is the presence of a LTR of 300–1,330 nucleotides at each proviral end. Initial experiments to identify the LTR sequences by hybridization of the 5′ and 3′ halves of the clones to each other failed to reveal any repeated sequences in the appropriate location. However, hybridization experiments with fragments of BaEV and HC-57 DNAs revealed that the 5′ portion of the BaEV LTR is homologous to a portion of

```
                    1725                          1745
          GluAsnProSerGlnPheLeuGluArgLeuCysGluSer
Human     TGAGAACCCCAGTCAGTTTTATGAGAGGCTCTGTGAGTCA
M-MuLV    ....TCT...TCGGCC..CCTA.....A..TAAG..AG.C
            . Ser .  . Ala . Tyr .  .  . Lys . Ala

                    1765                          1785
          TyrGlnLeuTyrThrProPheAspProGluAlaThrGlu
Human     TACCAGCTCTATACTCCATTTGATCCAGAGGCTACTGAA-
M-MuLV    ..T.GCAGG..C.....T.A...C..T....ACC.A.GGC
            . ArgArg .  .  . Tyr .  .  . AspProGly

                    1804                          1824
          AsnGlnLeuMetValAsnThrSerPheLeuSerGlnVal
Human     AATCAGCTCATGGTGAATACATCATTTCTTAGCCAGGTGC
M-MuLV    ..GA.A..A...-..TC..TG..T..CA..T.G...TCTG
          GlnGluThrAsn  . SerMet .  . IleTrp . Ser

                    1844                          1864
          GlnGlyAspIleLysTrpLysLeuGlnLysLeuGlu
Human     AGGGTGACATTAAGTGGAAACTTCAGAAACTGGAAGGTTT
M-MuLV    CCCCA......GG.A.A..GT.AGG..GGT.A....A...
          AlaPro .  . GlyArg . . GlyArg .  .

                    1881
Human     CGCAGGCATGAATTGCTACTCAGCT
M-MuLV    AAA.AA..A..--C...-TGG..A.
```

FIG. 4. Alignment of human sequences at site 2 with the *gag* p30 gene of M-MuLV. DNA sequences are presented as in Fig. 2. A portion of the inferred human amino acid sequence is shown above the DNA sequences, and the M-MuLV amino acid sequence is shown below the DNA sequences where it differs from the human amino acid sequence.

```
      TGTGTGCCACCCTACGAAGAAATAG AATATAGTGGTGCTTTCTATTAAAA
 50   TATTGCCTTTATAAAAAGCACCAAA GGGGGGAATGAAGTAGAAAATTAAA
100   AATAAGGAGTTTCTTTTTCCTATGA TAGAAAAGTTACTATTTAAAGTTTA
150   AAGGCCCCAAACAACCAACACCCCA CCAAGTGGGAAGTTTAAAAAGAATA
200   TTAACTGCCTGTCTGTGTACCATTA ACCATATCTTTCCCCGACATATTTT
250   GCAAGTTTGTAAATTCCTGTTTTTT TTTTCTGCTGCACAGCTGCAAGGTC
300   ACAAAACAGATAAGCATAGGCTGCA AAACATGTTTTCCCAAGATTAAGAC
350   ATGTCATAGAATGATTAACTGCCTT TGTTCTTGCTTCTGTAAGCTTGCTT
400   CCTGCATCATGTTTTCCGCACCCTT GCTTCCATAAAAGACGCACGCCCTC
450   TTTGTTTGGTGCTCAGACTTTCTGG ACACA
```

FIG. 5. The sequence of the apparent human LTR sequence. The underlined nucleotides are features characteristic of retroviral LTR sequences.

HC-57 DNA (Fig. 1, map positions 15.9–16.15). To determine whether a complete LTR is present, we determined the sequence of a 1.2-kb region (map positions 15.1–16.3). Shown in Fig. 5 is a 480-nucleotide sequence (region 6 of Fig. 1) which may represent such a LTR because it has several structural features in common with known LTR sequences (20), including both transcription initiation and transcription termination signals.

One structural feature, an inverted repeat with T-G at the 5′ end and C-A at the 3′ end, is common to the ends of all LTR sequences and is one of the characteristics by which retroviruses resemble transposable genetic elements. The first nine nucleotides of the human LTR, T-G-T-G-T-G-C-C-A, and the last eight nucleotides, T-G-G-A-C-A-C-A, would be a perfect inverted repeat with the addition of a C residue between the T-G-G and the A-C-A-C-A. These 9 nucleotides match 9 of 10 corresponding nucleotides in the LTR sequence of CPC-1 (21).

A LTR contains the transcription initiation sequences C-C-A-A-T and $\overset{G\text{-}T}{C\text{-}A}$-A-T-$\overset{T\text{-}T}{A\text{-}A}$-A-A-G, which are usually located 70–80 and 20–30 nucleotides, respectively, upstream of the guanosine

EUKARYOTIC	T	T	T	T	–	C	A	C	–	T G C A
HUMAN LTR	T	T	T	T	–	C	T	G	C	T G C A
RETROVIRAL	C/T	T	T	T	G	C	N	G/C	T	T G C A

FIG. 6. Comparison of transcription termination sequences. The consensus sequence for eukaryotic messenger RNAs is from Benoist *et al.* (22). The consensus sequence for retroviruses is derived from the published sequences (19, 21, 23, 24) of Moloney, AKR, BaEV, CPC-1, and MAC-1 viruses. The boxes enclose the common nucleotides of the proposed human termination sequence and the two consensus sequences.

at the RNA cap site. The human LTR has similar sequences, C-C-A-A-A (5 of 6 identities) and G-T-T-T-A-A-A-A-A (7 of 9 identities) beginning at nucleotides 157 and 187, respectively. The appropriate residue for the cap site would be the guanosine at either nucleotide 215 or 217. This site and the 5′ end of the LTR define the U3 region as ≈216 nucleotides in length compared to reported sizes of 227–1,200 nucleotides.

The cap site specifies the beginning of the R region, which normally terminates with the dinucleotide C-A at the poly(A) addition site. The consensus signal sequence for poly(A) addition, A-$\overset{A}{G}$-T-A-A-A, is located 20–30 nucleotides 5′ of the C-A. The presumptive human LTR contains the sequence T-G-T-A-A-A (five of six identities) beginning at nucleotide 258, which is 28 nucleotides 5′ of the C-A at nucleotides 286–287. At nucleotides 276–287 is the sequence T-T-T-T-C-T-G-C-T-G-C-A, which closely matches the eukaryotic and retroviral transcription termination signals (Fig. 6). This signal is always located within a few nucleotides of the poly(A) addition site. These sequences define an R region of 70 nucleotides, which is comparable to those of other retroviruses. The remaining 194 nucleotides of the proposed LTR would constitute the U5 region. This value is considerably larger than the largest reported U5 sequence (120 nucleotides). This identification of LTR regulatory signals is supported by the observation that the transcription termination site is separated from the polymerase sequence

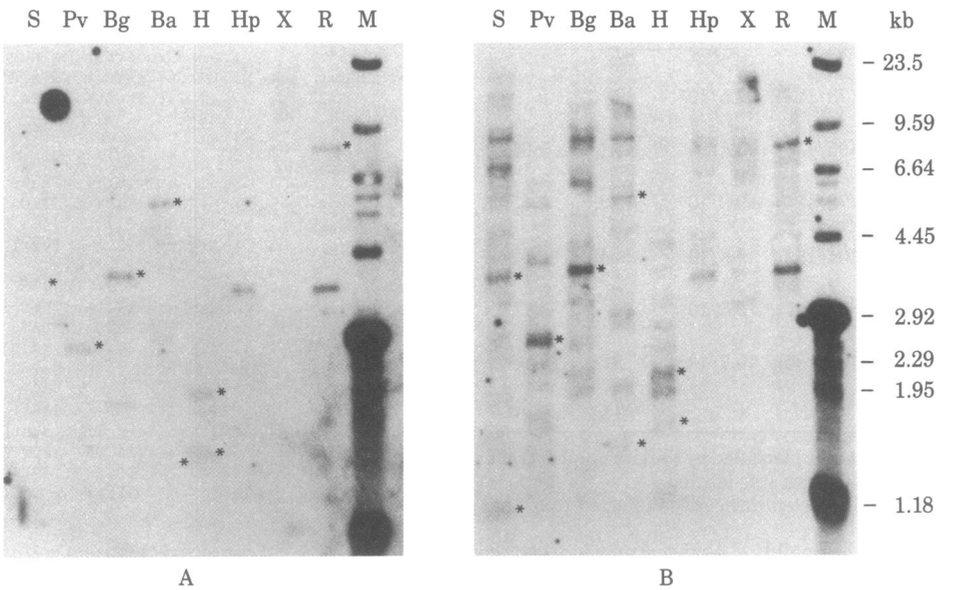

FIG. 7. Hybridization of cloned human DNA to restriction digests of human cellular DNA. Human DNA was digested by the indicated restriction enzymes (designated as in Fig. 1), electrophoresed in 0.7% agarose gels, transferred to nitrocellulose filters, and hybridized to labeled DNA containing the 2.3-kb *Eco*RI fragment of HC-20 (map position 12.6–14.9). (*A*) High-stringency hybridization. (*B*) Low-stringency hybridization. Asterisks to the right of individual bands identify the fragments expected from the restriction map of Fig. 1. The human DNA was extracted from a normal, adult kidney and is not the same as the fetal liver DNA used to construct the library. M, marker.

at site 5 by the same spacing as occurs in M-MuLV (Table 1).

Although its features implicate the 480-nucleotide sequence as a LTR, one aspect is presently unclear. The site of the hybridization between the BaEV LTR and the proposed human LTR is located at nucleotides 360–392. With the deletion of the T at 378 or 379, this sequence exactly matches 33 nucleotides of the BaEV sequence (24). The significance of this sequence is unclear because it occurs near the 5′ end of the U3 region of the BaEV LTR. However, the probability of obtaining 33 of 34 matching nucleotides is so small that this sequence conservation suggests some significance. If this sequence has a function in the U3 region, then the human LTR may extend several hundred nucleotides 3′ of the sequence in Fig. 5.

Related Sequences in Human DNA. Although the cloning results suggest that there are only a few copies in the human genome that are closely related to the 2.3-kb *Eco*RI fragment of HC-20, a more definitive assessment is provided by hybridization to restriction digests of human DNA. Low-stringency hybridization (Fig. 7*A*) reveals 10–20 bands in each digest. However the high-stringency result (Fig. 7*B*) shows only one or two bands per digest, many of which are predicted from the restriction map. Most digests contain an additional major band, which suggests the presence of closely related sequences that are divergent enough to have a different restriction map. The 8.0-kb *Eco*RI band would be expected if the *Eco*RI site that defines the 3′ end of CH2 is present in human DNA. Because the 3′ end of this fragment, as well as the 5.5-kb *Bam*HI fragment, must be in the flanking sequence, we assume that the intensity of these bands represents a single copy. This interpretation is consistent with the intensities of the other bands. Thus, the 3.7-kb *Eco*RI band would represent the second copy. We conclude that the human genome contains only two copies of sequence closely related to our clones but ≈10 copies of more distantly related sequence. There is no evidence of the specific multiple copy bands that are characteristic of the clone of Martin *et al.* (8) in similar blots with probes that span map positions 8.4–14.9. Thus, our clones represent an entirely different family of sequences.

DISCUSSION

We isolated several overlapping human clones containing significant DNA sequence homology with M-MuLV in the *gag* p15 and p30 genes and the polymerase gene. Furthermore, we have identified a possible 3′ LTR element at the same distance from the polymerase gene as in M-MuLV. Thus, the human sequences have the structure of an integrated retrovirus. However, we have been unable to identify a LTR at the 5′ end and, therefore, conclude that this provirus is incomplete. Although the spacing between the polymerase sequences and the 3′ LTR is the proper size to contain envelope genes, thus far we have been unable to recognize any substantial homology between this sequence and the Moloney virus *env* gene sequence. The absence of envelope gene homology is not unexpected because retroviral *env* genes are highly divergent. It appears that this provirus cannot code for a complete polymerase gene because the sequence at site 4a has no open reading frame.

From the similarity of the CH2 and HC-55 restriction maps, 18 of 19 sites, we deduce (25) that their overall sequence divergence is only 1%. This value is supported by the observation of only 6 nucleotide substitutions in 256 nucleotides of compared sequence and approximates the 1–2% divergence of the single copy sequences of chimpanzee and man. Because the two clones also cross-hybridize throughout the 3′ flanking sequence, we conclude that they are located at the same chromosomal site in their respective genomes. This common locus implies either that it is a highly preferred integration site or that the viral sequence was present at this site in the common ancestor of man and chimpanzee. However, it would require an extraordinary degree of preference for independent infections of man and chimpanzee to place the only closely related human viral copy at the same site as its chimpanzee homologue. Thus, we conclude that this is an ancient integration site and that this sequence has not moved in the sense of a transposable element during the last several million years. An alternative interpretation of the structure of the human sequences that would account for the presence of only a few incomplete ancient copies is that these sequences represent a protovirus (26), a precursor from which a retrovirus might be assembled.

Note Added in Proof. The human LTR is preceded by the purine tract, A-A-A-G-A-G-G-A-A, which is similar to those commonly preceding 3′ LTRs. Such tracts are the putative primers for (+)-strand DNA synthesis (27).

We thank Edward Birkenmeier for his role in the cloning of CH2; George Searfoss, Steve Kerby, and Marilyn Powers for their technical assistance; and Karen McNitt for her assistance with the computer analysis. Part of the work upon which this publication is based was performed pursuant to Contract NO1-CO-75380 with the National Cancer Institute.

1. Blair, D. G., Oskarsson, M., Wood, T. G., McClements, W. L., Fischinger, P. J. & Vande Woude, G. F. (1980) *Science* **207**, 1222–1224.
2. Wong-Staal, F., Dalla-Favera, R., Franchini, G., Gelmann, E. P. & Gallo, R. C. (1981) *Science* **213**, 226–228.
3. Neel, B., Hayward, W., Robinson, H. & Astrin, S. (1981) *Cell* **23**, 323–334.
4. Varmus, H., Quintrell, N. & Ortiz, S. (1981) *Cell* **25**, 23–26.
5. Chattopadhyay, S. K., Cloyd, M. W., Linemeyer, D. L., Landers, M. R., Rands, E. & Lowy, D. R. (1982) *Nature (London)* **295**, 25–31.
6. Poiesz, B. J., Ruscetti, F. W., Gazdar, A. F., Bunn, P. A., Minna, J. D. & Gallo, R. C. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 7415–7419.
7. Reitz, M. J., Poiesz, B. J., Ruscetti, F. W. & Gallo, R. C. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 1887–1891.
8. Martin, M. A., Bryan, T., Rasheed, S. & Khan, A. S. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 4892–4896.
9. Benton, W. D. & Davis, R. W. (1977) *Science* **196**, 180–182.
10. Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503–517.
11. Birkenmeier, E., Bonner, T. I., Reynolds, K., Searfoss, G. H. & Todaro, G. J. (1982) *J. Virol.* **41**, 842–854.
12. Dayhoff, M. O. (1976) *Atlas of Protein Sequence and Structure*, Vol. 5, suppl. 2. (National Biomedical Research Foundation, Washington, DC).
13. Shen, S., Slightom, J. L. & Smithies, O. (1981) *Cell* **26**, 191–203.
14. Cleary, M. L., Schon, E. A. & Lingrel, J. B. (1981) *Cell* **26**, 181–190.
15. Van Ooyen, van den Berg, J., Mantei, N. & Weissmann, C. (1979) *Science* **206**, 337–344.
16. Lawn, R. M., Fritsch, E. M., Parker, R. C., Blake, G. & Maniatis, T. (1978) *Cell* **15**, 1157–1174.
17. Bonner, T. I., Birkenmeier, E. H., Gonda, M. A., Mark, G. E., Searfoss, G. H. & Todaro, G. J. (1982) *J. Virol.* **43**, in press.
18. Maxam, A. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
19. Shinnick, T. M., Lerner, R. A. & Sutcliffe, J. G. (1981) *Nature (London)* **293**, 543–548.
20. Temin, H. M. (1981) *Cell* **27**, 1–3.
21. Lovinger, G. G., Mark, G., Todaro, G. J. & Schochetman, G. (1981) *J. Virol.* **39**, 238–245.
22. Benoist, C., O'Hare, K., Breathnach, R. & Chambon, P. (1980) *Nucleic Acids Res.* **8**, 127–142.
23. Van Beveren, C., Rands, E., Chattopadhyay, S. K., Lowy, D. R. & Verma, I. R. (1982) *J. Virol.* **41**, 542–556.
24. Tamura, T., Noda, M. & Takano, T. (1981) *Nucleic Acids Res.* **9**, 6615–6626.
25. Nei, M. & Li, W.-H. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5269–5273.
26. Temin, H. M. (1974) *Cancer Res.* **34**, 2835–2841.
27. Varmus, H. E. (1982) *Science* **216**, 812–820.