S71 Is a Phylogenetically Distinct Human Endogenous Retroviral Element with Structural and Sequence Homology to Simian Sarcoma Virus (SSV)

THOMAS WERNER,*¹ RUTH BRACK-WERNER,† CHRISTINE LEIB-MÖSCH, HORST BACKHAUS,§ VOLKER ERFLE,† AND RÜDIGER HEHLMANN¶

*GSF-Institut für Säugetiergenetik, †GSF-Abteilung für Molekulare Zellpathologie, and §GSF-Institut für Strahlenbiologie, Ingolstädter Landstrasse 1, D-8042 Neuherberg, ‡Medizinische Poliklinik der Universität, D-8000 München, and ¶III. Medizinische Klinik Mannheim der Universität Heidelberg, 6800 Mannheim, Federal Republic of Germany

Received May 2, 1989; accepted September 5, 1989

Human endogenous retroviral element S71 had previously been shown to contain gag- and pol-related regions and a 3' LTR-like sequence. The nucleotide sequence of S71 was determined and compared with the corresponding regions of SSV and its helper virus SSAV. The 1.48-kb S71 gag region consists of matrix protein p15 (MA)-, capsid protein p30 (CA)-, and nucleocapsid protein p10 (NC)-related sections and the 1.82-kb pol region of tether, RNase H (RH), and endonuclease/integrase (IN) sections. The S71 nucleotide sequence contains a 167 amino acid open reading frame encompassing MA. The boundaries of the S71 element are delimited by direct repeats and the entire element is 5.4 kb long. Similarity between S71 and the v-sis-bearing, defective SSV provirus also covers overall structural organization, including the presence of presumably nonretroviral sequences. Both the gag and the pol regions of S71 contain sequences highly conserved in numerous retroviruses. Phylogenetic analysis with conserved CA, RH, and IN sequences showed that of all other (C-type) human retroviral elements available for comparison, S71 is most closely related to infectious primate and murine retroviruses. This suggests that S71 represents a phylogenetic subgroup of its own. In addition we identified short ranges of conserved amino acid sequences within C-type retroviral gag and pol genes sufficient for phylogenetic analysis. Use of these may facilitate large-scale phylogenetic evaluation of C-type retroviral elements. © 1990 Academic Press, Inc.

INTRODUCTION

About 10% of the human genome is thought to have arisen by reverse transcriptase-mediated processes. Sequence information generated in this manner has been given the general term retroposon (Temin, 1985). Retroposons constitute a relatively large, heterogeneous group which can be divided into nonviral and viral sequences (Weiner et al., 1986). The group of nonviral sequences is composed of LINES (long interspersed nuclear sequences; for review, see Skowronski and Singer, 1986), SINES (short interspersed nuclear sequences), processed genes, and pseudogenes. The group of viral retroposons consists of DNA sequences related to infectious retroviruses. On the basis of copy number and length, one can calculate that these retroviral elements make up at least 0.1% of the human genome. Although some of these elements show extensive structural similarity to full-length infectious animal proviruses, the presence of these elements is not the result of a singular somatic infectious process limited to one generation and to a specific cell type. Rather, they are an endogenous part of the human genome and are transmitted from one generation to the next in a stable Mendelian fashion. Sequence similarity with C-type mammalian retroviruses on the one hand and A-, B-, and D-type retroviruses on the other allows division of human endogenous retroviral elements into two groups designated class I and class II, respectively (Callahan, 1988).

Retroviral polymerase genes consist of at least three regions coding for separate activities: the reverse transcriptase (RT), RH, and IN. The approximate location of these sections within the polymerase gene has been mapped with the help of bacterial expression clones containing various regions of retroviral polymerase genes and subsequent testing for retrovirus-specific activities (Hansen et al., 1988; Hizi and Hughes, 1988; Levin et al., 1988). These experimental data agree with previous computer analysis predicting the structural organization of retroviral polymerase genes (Johnson et al., 1986). The major region of homology between various retroviruses is located within their pol genes (Chiu et al., 1984). Multiple alignments of retroviral pol amino acid sequences show that the RNase H domain and the endonuclease region of retroviral pol genes contain highly conserved sequences which may be essential for their functionality (Johnson et al., 1986; Doolittle et al., 1989). Therefore sequences from these sections of the pol gene should be suitable for phylogenetic analysis of class I elements by comparison with each other as well with C-type animal retroviruses. As-

¹ To whom requests for reprints should be addressed.

suming that these results reflect actual evolutionary relationships, one would expect congruent data from highly conserved sequences located in other C-type retroviral genes. Since numerous class I retroviral elements also contain gag-related regions (Brack–Werner *et al.*, 1989a,b), we decided to look for the presence of highly conserved sequences suitable for phylogenetic analysis in C-type retroviral gag genes.

Molecular clone S71 contains a class I element isolated by low stringency hybridization with recombinant probes derived from the C-type simian sarcoma-associated virus (SSAV). Hybridization analysis with SSAV probes and limited sequence data showed that S71 is an incomplete proviral element consisting of gag-polrelated sequences (Leib-Mösch et al., 1986). At its 3' terminus it contains a 535-bp sequence with features characteristic of retroviral LTRs (Brack-Werner et al., 1989a). The putative control regions of the S71 LTRlike sequence show a higher degree of sequence similarity to infectious murine and primate proviral LTRs than to other human endogenous retroviral LTRs. The detailed structural analysis of S71 reported here shows that numerous other features of C-type retroviruses are also conserved in the S71 element. To determine whether our initial observation of closer relationship of S71 to primate murine and infectious proviruses is reflected by the complete S71 element we decided to utilize conserved sequences of the gag and pol regions for phylogenetic analysis. Since, in many cases, only fragmentary sequence data are available, phylogenetic analysis of endogenous class I elements was carried out with short conserved sequence stretches (40-50 amino acids) and compared with longer sequence stretches (150-230 amino acids). Our results indicate that S71 is more closely related to the infectious murine and primate proviruses AKV, BaEV, and SSV than to the other class I human endogenous retroviral sequences ERV1, ERV3, 4-1. Therefore S71 may represent a phylogenetically distinct subgroup of class I endogenous retroviral elements.

MATERIALS AND METHODS

Nomenclature

Retroviral proteins are designated using the nomenclature proposed by Leis *et al.* (1988). Subregions of proteins are referred to by an additional lower case letter in parenthesis (e.g., CA(s), CA short region).

Oligonucleotide preparation

Specific sequencing primers (oligonucleotides) were synthesized on an Applied Biosystems Model 381 A DNA synthesizer and purified either by electrophoresis in 14% polyacrylamide gels containing 8 mol/liter urea and subsequent desalting as specified in Brack–Werner *et al.* (1989a) or with OPC cartridges (Applied Biosystems, England).

DNA sequence analysis

S71 subclones were generated by cloning restricted S71 DNA fragments in pUC vectors (Vieira and Messing, 1982). Dideoxy sequencing of double-stranded plasmid DNA was carried out as described by Chen and Seeburg (1985) using either Klenow fragment (Pharmacia, Sweden) or the modified T7 DNA polymerase (Tabor and Richardson, 1987) Sequenase (United States Biochemical Corporation). Both strands were sequenced as shown in Fig. 1.

Computer alignments and construction of phylogenetic trees

Dot matrix analysis (Fig. 3) was carried out with Mac-Gene Plus software (Applied Genetic Technology, Inc.) on a Macintosh Plus computer. Sequence alignments were performed with the GENALIGN program (GEN-ALIGN is a copyrighted software product of IntelliGenetics, Inc.; the program was developed by H. Martinez of the University of California at San Francisco) and the program package for phylogenetic analysis provided by R. F. Doolittle (Feng and Doolittle, 1987; McClure et al., 1988). The phylogenetic trees shown in Fig. 6 were obtained with the latter program package since it provides a full analysis. Sequences were taken from the EMBL database release 10. The particular versions of sequences employed are human endogenous retroviral elements 4-1 (Repaske et al., 1985), ERV3 (O'Connell et al., 1984), and ERV1 (Bonner et al., 1982); baboon endogenous virus (BaEV; Kato et al., 1987), SSV (Devare et al., 1983), SSAV (Brack-Werner et al., 1989c), AKV (Etzerodt et al., 1984), human immunodeficiency virus (HIV-1; Ratner et al., 1985), mouse mammary tumor virus (MMTV; Moore et al., 1987), squirrel monkey virus (SMRV; Chiu et al., 1984), Rous sarcoma virus (RSV; Schwartz et al., 1983), and COPIA (Mount and Rubin, 1985). The sequence of the chimpanzee endogenous virus is to our knowledge published only in the database (accession No. K02915).

RESULTS

S71 sequences related to the SSV gag gene

The gag gene of SSV codes for a 65-kDa gag precursor protein (pr65^{geg}) which is processed to yield MA, p12^{geg}, CA, and NC (Teich, 1984). Within the gag gene the sequences coding for these proteins are arranged 5'-MA-p12^{geg}-CA-NC-3' (Devare *et al.*, 1983). In the S71 element sequences related to the SSV gag encoding gene extend from position 498 to position 1975 (Fig.



Fig. 1. Nucleotide sequencing strategy of S71 and organization of retrovirus related sequences. A physical map marking the positions of restriction sites relevant in the generation of S71 subclones is depicted at the top. H, *Hind*III; K, *Kpn*I; P, *Pst*I; Pu, *Pvu*II; S, *Sac*I. Arrows indicate the extent and direction of nucleotide sequences obtained from one or more sequencing experiments with an individual subclone of S71. Parallel arrows covering the same sequence range indicate sequencing runs carried out with various overlapping subclones. The range of retrovirus-related sequences in S71 was determined by alignment with the corresponding SSV/SSAV sequences (see text). Vertical solid lines mark the boundaries of MA/p12^{gag}, CA, and NC sections of the gag region and the tether, RH, and IN sections of the pol region. Stippled regions in the three reading frames deduced from the S71 nucleotide sequence indicate the relative location of SSV gag (light stippled) and SSV/SSAV pol (dark stippled) amino acid sequences. The positions of stop codons in each frame are marked by vertical lines.

2). Translation of this region of S71 shows that the deduced amino acid sequences contain extended regions of similarity with the SSV MA, CA, and NC sections and less similarity in the p12^{gag} corresponding region (Figs. 2 and 3). The largely uninterrupted diagonal in the computer-generated dot matrix analysis indicates extensive collinearity of both sequences.

The SSV gag-related amino acid sequences in S71 are dispersed among various reading frames (Fig. 1), indicating that multiple frame shifts preclude the synthesis of the full-length gag precursor protein. However, a possible open reading frame extending from position 498 to 1004 (168 AA), marks the beginning of the S71 gag-related region and encompasses the entire MA-related section (Fig. 2). This reading frame begins with a methionine codon which matches in 7 out of 9 positions with the consensus sequence CC A/G CCAUG(G) established by Kozak (1984).

Oroszlan *et al.* (1977, 1981) found the NH₂-terminal region of the major internal virion protein of C-type retroviruses (CA) to contain the following highly conserved sequence: $PLR(X)_{7-15}YWPES(X)SDLYNWK$. The deduced amino acid sequence of S71 between

positions 1041 and 1135 is identical to the above conserved sequence in 11 out of 15 positions (67%, underlined above), indicating that this region of S71 constitutes the 5' terminus of the CA-related sequences. The largely collinear alignment of the S71 CA region with the corresponding CA region of SSV (Fig. 3) locates the 3' terminus of S71 CA-related sequences around position 1804 (Fig. 2). Sequences related to the NC coding region of the SSV gag gene extend from position 1805 to 1975 in S71. The amino acid sequences of C-type retroviral NC proteins contain a highly conserved, cystein-rich structure implicated in nucleic acid binding (Covey, 1986; Fütterer and Hohn, 1987). In the S71 NC this $CX_2CX_4HX_4C$ nucleic acid binding motif is fully conserved (positions 1913–1954).

The deduced amino acid sequence of the S71 gag region shows an overall identity of about 40% with the amino acid sequence of the SSV gag gene (underlined amino acids in Fig. 2). From the computer-generated dot matrix analysis it is evident that similarity is lowest in the p12^{gag} region (Fig. 3), which is also reflected in the relatively large number of gaps obtained in a Needleman–Wunsch alignment of these sequences (data

Kpn I									
GGTACCGCCC	AGTCATCC	IG GCAACCCCG	I GCTGCTCAGC	AGGGCTCCTC	CCAGCCTGAA	AACATCTGAG	TGGCCCCTTT	CCTCCTCATC	90
	ACCCCGCA	CA TCCCGTTTT	C CTGTGCCACA	GCAAGTCCAG	CGCCTCCAAG	ACTTGGCTCC	GCTCTCCCTC	CTAAAATCCT	180
TAAAAGAAAG	GGCAAAGT	rt gaacttttt	I CCTTCAAGTC	GTGGAGACGC	САААААТАТТ	ТАБАСТАТАА	GTCAGAGAGG	AGAGGGGGAT	270
CACGTAGGTC	CCACTAGO	CT CGCACCCAT	C TCTTGTCCTC	TCCCTAAATC	TTGGAGCTTA	AGGAAACAGA	CCTTATGTGG	CAAGAAGCGC	360
TGGCTATAGC	TGTTTTCC	IA CTTCTTTTG	G TTATGATGCI	TCTATTCTTC	CGATACTCCA	GCCCCTCCAG	GTCATGAATT	TCTCTGTCCA	450
TGCTGGGTTT	ААТАТСТС	IG CTCAAACIT	I GTTAAACTGC	CTCCAGAATG	GGAAACTCTT G N S S	-> CTTCCCAGTC SQS	TCATAAAGAT H K D	TGGAGCCCTC WSPL	540
TCCAATGTAT Q C M	GTTACAAA LQN	TTCTCTCTA	G GCTTCTCCGA G F S E	. GGATTATGGG DYG	GTCCGCCTTA VRLK	AAAAAGGCAA KGK	L W T	CTCTGTGGAG L C <u>G V</u>	630
TAGAATGGCC E W P	AAAGTTTG K F G	A GCCGGGTCA	C TGAACCTCGC	AATTGTTCAG IVQ SacI	GCTGTGTGGC	GGGTTGTTGC VVA	TGGAACTCCT G T P	GGTCACCCTG G H P D	720
ATCAGTTTCC	CTACATTG	AT CAATGGCTG	AGTTTGGTCCG S <u>L</u> VR	GAGCTCTCAT	CCATGGCTCC PWLH	ACTCATGCGC S C A	CATTCCTAAT I P N	CCTACCTCCA P T <u>S</u> K	810
AGGTCATTTT VIL	GAGCCAGA S <u>Q</u> T	CC TCACTTTCG	C CTCGACCCTC P R P S	AGCCGGCTCG A G S	GCTCCTCCTG	TATTGCCTCC L P P	TTCTGAAGAA S E E	GAGGAAAGTC E E S L	900
TCCCTCACCC	AGTTCTGC	CG CCTTATAAC	C CTCCTGCTCC	CTTAGAATCT L E S	TCCCTTGTCT	CCTCGACTAC	ATCCCCTGTG S P V	GGCTCTCTGC G S L P	990
CTATTGCCTC	CTGATTGA X L R	GG CCACAGCAGA	G AGGAGGTAGC E E V A	P L L	CTGCTGAGAG L L R E	AGGCACAAGT A Q V	CCCTGCGGGT PAG	GATGAGTGCT D E C S	1080
CAGCTCCATT A P T	CTTGGTTT. L V Y	AT GTCCCCTTT	T CTACTTCTGA S T S D	CCTGTGCAAC	GGAAGGCTCA K A H	TAATCCCTCC N P S	TTTTCTGAAA FSEK	AGCCCCAGGT	1170
CTTGACCTCA	CTGATGGA	ST CGGTGCTCTC S V L W	GACCCATCAA T <u>HQ</u>	CCCACCTGGG P T W D	ATGACTGTCA D C Q	ACAACTCCTT Q L L	TTAACCCTCT L T L F	TCACCTCTGA T S E	1260
AGAGAGGGAT ERD	CGTATCCG	AA GAGAAGCCAG R <u>EA</u> R	GAAAGTATTTC KYF	CTTACATTAG	CCGGTAGACC	GGAGGGGGAA E G E	GCCCAAAACC	TCCTTGAGGA L E E	1350
GGTTTTTCCC	TCTACCCGG	CC TGATTGAGAS	FCCGAACTCCT P <u>N</u> SS	CAGGTGGGAA G <u>G</u> K	GAGAGCTTTG R A L	GATAATTTTC DNFH	ACCGTTATCT	CCTTGCGGGT	1440
ATCAAGGGAG I <u>K</u> GA	CCGCTCGA	AA ACCATGAATO	L ST	AACTGAAGTT T <u>EV</u>	GTCCAGGGGC VQGP	CTGGTGAGTA G E X	ACCTGGAGCA PGA	TTTTTAGAAT FLEC	1530
GCCTCCAGGA	GGCCTATC	T Y T I	CTTTTGACCC	AGCGGCTCCC A A P	GAGAAGAGCC E K S R	GTGTTATTAA V I N	TTTGGCATTT L <u>A F</u>	GTGGCTCAGG V A <u>Q</u> -	1620
CGCCTCTGAT -A S D	ATTAGAAAA IRK	A AATTACAAA KLQK	ACTGGAAGGA	TTTGCTGGAA F A G M	TGAACATTAG N I S	CCAGCTTTTA Q L L	GAAGTAGCCC E V <u>A</u> Q	AGAGAATTTT R I F	1710
TGACAGTCAA DSQ	GAGTTCGAG E F E	GA AACAAAAACA KQKQ	A GGCAGCTGAA A A E	AAGGCTGCTG K A A D	ATGAAACATC E T S	CAAAAGACAA KRQ	CCGAAAATCT PKIL	TAGTGGTCGC VVA	1800

Fig. 2. Nucleotide sequence of human retroviral element S71. The relative positions of restriction sites shown in the schematic in Fig. 1 are indicated. The S71 retroviral element is delimited by direct 7-bp repeats (arrows). The sequence of the 3' LTR-like region extending from position 4975 to 5509 has been published previously (Brack–Werner *et al.*, 1989b). The S71 nucleotide sequence was translated and retrovirus-related amino acid sequences were identified by a Needleman-Wunsch alignment with SSV/SSAV and AKV protein sequences. The SSV/SSAV-related deduced amino acid sequence of S71 is depicted below the nucleotide sequence and amino acid residues in common with SSV/SSAV are underlined. Interruption of underlining between adjacent amino acid residues marks gaps in the S71 amino acid sequence required for maximum similarity with the SSV/SSAV protein sequence. Termination codons are marked with an X and dashes indicate frame shifts. The borders of the MA, CA, and NC gag regions were inferred from alignment with the gag gene of SSV. Additional confirmation of the 5' border of the S71 CA

r-p 10 ->	
CATCCOGGAA GCCAGAAAGG AGGGGCCCCC ATCACAGAAC ACTAGCCAGG GGACCCCGGT TCCACAACAG AAAGGCCAGA AAAGTGAGTA I R E A R K E G P P S Q N T S Q G T P V P Q Q K G Q K S E X	1890
A S L Q K N Q C T Y C K Q I G H W K K E C P F K P E G K I M	1980
	0505
PIKARGETRKQKGLKGPAPPHLWVFLIRWV	2070
	101.00
KRVRKED SRGKVIEKELWPRDQRLAYRGPA	2160
	2250
LALVSEFHQYLLITISTISERGMWQDYRVM	2250
	2240
V G R G S A G T H V S K D L C V I N K F E E R C C A L M R T	2340
CHACCOLLS WHILE MORE ACTIVITY ALL ALL MORE COLLEGE ACTIVITY ALL ACTIVITY COLLEGE COLLEGE COMPANY	2420
GIAGGCAGA ITTAIGITIG ACTITACACA AACATCICGS IGCATTAAAG AGCAGTATIG CCGCCAGCAT GITTCACCIC CAGCCATAAG X	2450
እርእናሚመምመቁ የተመልመረም እና መል ልልመልናልልና የሚያምንልመመቁ የማማመልናልናም የእናእናልመምርና <u>እመምርር</u> እርር እርር እር እና እርር እሸር የ	2520
AGGITTITI CLATCICAG TANATAGANC GIALGATITI GITTACACI GAGACATICC ATTCCCAGGS ACCGAGGAGG AGACGGATGC	2320
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	2610
	2010
	2700
GENTITUU TIULAUGAG GULATATUT AGGUGIAL ATGGGAGAA AGUTIGGALA ATACATGGUT TIULIGGULA GAGULUUTG	2700
	2700
TOSCUTUTG CAATGUATTG TGTCULTOGG TACTUGAGAT TAGAGAATGG CAATGACITA CUAAGUATAC TGUUTUAAA CAUATTITUA	2790
	2990
ACAAAGCACA ICCIGCACAG CCCIAAAICC AIIAAACCII GAGICAACAI AGCACAIGIC ICIGCAAGCA CAGGIIGOG GCIAGGGIIA	2000
ለት የመጠት ትላይ በሚያመረስ እና በረት እንደ አለት አመ መግናመናመመስ የመ እና አርት አለት አመር የሰጥ መግለ መግለ የመስጥ የሚያመር የሚያመር የሚያመር እና አርት እና አ	2970
CAGATTAACA GCAILTCAAG GCAGAAGAAT TICICTTAGT ACAGAACAAA AIGGAGITIC TIAIGICIAC TICITICIAC AIAGAGIAAC	2370
	30.60
AGICIGATET ETETETATTI TECCARAAA GOTETEAET ETECEGORG CAGAGAARTE TGATGATTAR ARGGEEGOG GETECETETE	5000
	31.50
C P V I Y L S K R L D P V A	- 5150
	3240
SRWTSCLWAIAAGIGICGUA AACAAGUGA AAAAGUGA AAGUGA A	5240
	- 2220
L T A P R A I E T L L Q S A S G K W M S N A R I L Q Y Q S L	
	3420
L L D W P R L T F S P T R C L N P A T L L P D P D F T T P V	5420
	- 3510
H D C Q E L X E T T E T V R P D L Q D V P L K E V D A T V F	2210
Ряс I Расколос сталинские састасских состоется состаться телевосска талистоса астолосова	3600
TDSSSLL LKQGVRKAGA VTMETDKLQA	5000

region was obtained by the presence of a sequence highly conserved in the NH₂ terminal region of C-type retroviral gag proteins (see Results). Assignment of the S71 pol amino acid sequences to the tether, RH, and IN sections of the polymerase gene is based on the data of Johnson *et al.* (1986). In the RH region, 11 highly conserved amino acid residues (Levin *et al.*, 1988) and their relative locations in the S71 sequence are marked with asterisks. In the IN region, asterisks mark the relative locations of conserved amino acid residues (Levin *et al.*, 1988).

TGCC	AGC	CAGG	TAC	Стс	AGCA	CA	AAA	AGC	rG	AATT	GGT	IGC	TCT	CAT	TCAG	GCO	сст	CCGI	AC	GGGT	AAG	GAC	ААА	TGT	ATTA	AC	ATT	тас	AC	3690
Р	A	G	Т	s	A	Q	к	Α	Ê	L	v	Å	L	. I	Q	À	L	R	R	V D	R K	Т D	N K	c	L I N	T	I	Ϋ́	T T	
TGAC	AGC	CAGG	TAT	GCI	TTTT	GC	TAC	TGT	GC	GTGT	ACA	TGG	AGC	CAT	CTAC	CA	AGT	GCG'	ΓG	GGCT	АСТ	CAC	CTC	AGC	AGGA	AA	.GGC	TAT	TA	3780
т D	A S	G R	M Y	L A	F	A	T	v	R	v	н	G	<u>A</u>	I	Y.	Q	v	R	G	L	Ŀ,	T	S	A	G	ĸ	. A	1	ĸ	
AAA	CT	AGA	AGA	AAT	TTTG	GC	сст	GCT	TG	AAGC	TGT	TTG	ССТ	ACC	TCAA	CA	GGT	GGC'	ľG	ТААТ	TCA	CTG	CAA	AGG	AÇAT	CA	ААА	AGA	AG	3870
N	х	E	E	I	L	A	L	L	E	A	v	с	L	Р	Q	Q	<u>v</u>	A	v	I	н	с	ĸ	G	Ĥ	0	ĸ	E	D	
ACAC	GGC	CGT	TGC	CCA	TGGT	AA	CCA	AAG	AG	CAGA	L CTC	'st IGC	AGC	CTG	GGGG	cci	AGC	TCA	AC	TGCC	AGT	cgc	GCC	TCC		CT	GCT	GCC	SC 1 TG	3960
T	A	v	A	н	G	* N	Q	R	A	Ď	s	A	A	W	G	Ρ	A	Q	L	Ρ	<u>v</u>	A	Ρ	Р	T	L	L	P	A	
CAGI	GTC	CTT	TCC	GCA	ACCT	GA	ርጉጥ	GTC	AG	ATCA	ccc	AGA	АТА	TTC	CCCA	GA	GGA	GGA	AA	AACA	GGC	TTC	GGA	TCT	TCAG	GC	CAG	таа	АА	4050
V	S	F	P	Q	P	D	L	S	D	Н	P	Е	Y	s	Р	E	Е	E	ĸ	0	A	s	D	L	Q	A	s	ĸ	N	
ATC	GGJ	AAGG	AGG	AGT	AAAA	СТ	GGC	CCA	GC	TTCT	AAG	GAG	CCG	TTT	CAAG	AT	ccc	CAA	cc	TTCA	GGA	CTT	AGT	TAA	CCAA	GC	AGC	TCT	ст	4140
Q	E	G	G	v	ĸ	L	A	Q	L	L	R	S	R	F	к	I	P	N	L	Q	D	L	V	N	Q	A	A	L	W	
CCTY	ምእስ	TOOT	TTC	тс <b>с</b>	0030	ст		CAC	ጥአ	3CC 3	100	TCC	ጥአአ	<b>»c</b> c	CACC	ምር	100	CCA	~~	GCCT	cca	666	202	CTC	2002	66	363	220	ርጥ	4230
Č.	T	V	c t	A	Q	v	N	T	ĸ	Q	G	P	K	P	S	s	G	D	R	L	Q	G	D	s	P	G	E	R		4250
~~~~			202	אא	ההההי	~~	303	~~~	~~	CACC	<u>ста</u>	~~~~	አጥአ		ጥጥ	ст	እርጥ	ACT		2020	ርጥጥ	ጥጥ	TCC	እጥል	ርእርጥ	GN		ልጥጥ	mc.	4320
E	I	Ť	E	I	K	P	H	W	A	G	Y	K	Y		L	¥.	L	V	D	T	F	ŝ	G	X	T	E	<u>A</u>	F	A	4520
			~~~		-		~~~	~~~			0.000							~~~					~~~					<b>.</b>	mo	
T	K	N	E	GAC T	A	AC T	T	V	V	K	F	S	ACT L	N	E	AD I	I	P	Q	H	G	GCT L	P	TAC T	A	M	GGG	GTC S	D D	4410
ATA	TAC	GATC	GGC	CTT	CACC	TO	GTC	CAT	AG	CTCA	GTC.	AGT	CAG	TAA	GGCA	TL	AAA	CAT	rC	AATG	GAA	GCT	CCG	TTG	TGCC	ТА	TCG	ACC	cc	4500
N	R	S	A	F	Т	S	s	Ι	A	Q	s	v	S	ĸ	A	Ļ	N	Ī	Q	W	K	L,	R	с	A	Y	R	Ρ	Q	
AGAG	кт	TGG	ATG	GGT	'AGAA	CA	CAT	GAA'	гc	ACAC	CCT	AAA	ААА	TAC	TGTT	AC	AAA	ATT	GA	TCTT	AGA	GAC	CGG	TAA	ааат	CA	GGT	AAG	AC	4590
S	s	G	W	*	Ē	H	м	Ň	н	T	L	ĸ	N	T	v	T	K	L	I	L	E	T	G	ĸ	N	Q	V	R	L	
TCCI	TCC	TTT	AAC	CCT	TCTT	AA	AGT	AAG	АТ	GCAT	TCC	гта	CCG	GGC	TGGG	TT	TTC.	ACC	гт	TTGA	ААТ	CAC	GTA	TAG	GAGG	CG	TCC	GCC	TA	4680
L	P	L	Т	L	L	K	v	R	с	I	P	Y	R	A	G	F	s	₽	F	E	I	T	Y	R	R	R	P	Р	I	
TCTT	GCC	TAA	GCT	AAA	GGAT	AC	CCG	ፐፐፐ	AG	CAGA	AAT	стс	AGA	AGC	таат	ጉጥ	АТТ	ACA	ЭT	ACCT	ACA	GTC	тст	CCA	ACAG	GT	ACG	AGA	ТА	4770
L	P	K	L	к	D	Т	R	L	A	E	I	S	E	A	N	L	L	Q	Ŷ	L	Q	s	L	Q	Q	v	R	D	I	
100 M 1	~~,		201	m c m	~~~	~	Sac		<b></b>	~~~~			moo				~~~	~~~~	~~	CCC1	~~~	~~~~~	~~~~			~	~~	<b>~</b> ~~	~	4960
I	Q	P	L	V	W	G	AGC.	H	P	S	P	V	P	D	Q	T	G	P	C	H	S	F	P	P	G	D	L	V	L	4860
								•			•	-																		
TAAP	AAG	STTC	CAG	GTT	AAAG	TT	TAA	AAA/	AA	AAAA 	AAA	GTT	CCA	GAA	AGAA	GGi	ACT	CAC	rc	CTGC	TTA	GAA	AGG	ACC	TCAT	AC	TGT	CAT	сс	4950
ĸ	s	5	R	г	ĸ	F.	ĸ	ĸ.	ĸ	ĸ	ĸ	F.	Q	ĸ	E	G	i. re	T' peat	P	A	X									
TCAC	CAT	IGCC	GAC	AGC	TCTG	GA	AG 1	TG.	L	TR s	equ	ence	e 53	5 bj	p(		CTT	ACC	c											5516
													Fi	<u> </u>	<u> </u>	ont	inu	ed												

not shown). Within the S71 gag region, the CA sequences show an amino acid identity of 47% (119/253 amino acids), indicating that this region is most highly conserved.

#### S71 sequences related to the SSAV pol gene

We had previously identified a 1.2-kb SSAV pol probe which yields a relatively strong hybridization signal with S71 (Leib–Mösch *et al.*, 1986). Sequence analysis of this fragment and part of the adjacent (3') fragment procuring an overlap with the published SSV sequence was carried out (Fig. 4; Brack–Werner *et al.*, 1989c). This allowed generation of a 2205 nucleotide SSAV/ SSV composite pol sequence structured 5'-tether-RH-IN-3' (Fig. 4) for comparison with the S71 pol region.

In the S71 element a section approximately 1800 nucleotides in length (Fig. 2, position 3111–4928) shows extended similarity with the SSV/SSAV sequence in dot-matrix analysis (Fig. 3). Under the conditions chosen for this analysis (see legend to Fig. 3) similarity between both sequences continues through several stretches in a collinear fashion and seems to taper off towards the end of both sequences. S71 sequences between nucleotide 1975 and 3111 or downstream of nucleotide 4928 showed no similarity to the SSV/SSAV

Kon I



Fig. 3. Comparison of the DNA and deduced amino acid sequences of the S71 gag and pol regions with the corresponding SSAV/SSV sequences. The diagonal lines shows the extent of similarity and collinearity of the S71 gag and pol regions with the SSV gag gene and the composite SSAV/SSV pol sequence. No similarity was observed beyond these regions. Dots in both protein and DNA sequence comparisons signify a minimal match of 56% in a window size of 20 residues. gag region: *X* axis, S71 nucleotide position 498–1978 (DNA); 491 amino acid residues (protein). *Y* axis, SSV nucleotide positions 1398–2936 (Devare *et al.*, 1983); 513 amino acid residues. pol region: *X* axis, S71 nucleotide position 3111–4928 (DNA); 606 amino acid residues (protein). *Y* axis, SSAV/SSV composite pol sequence (Brack-Werner *et al.*, 1989c) nucleotide position 239–2044; 602 amino acid residues (encompasses SSV nucleotide position 2903–3785; 294 amino acid residues).

sequence or any other retroviral polymerase gene available for comparison.

#### Organization of S71 pol sequences

The deduced amino acid sequence of the S71 pol region showing a high degree of similarity with the SSAV/SSV pol protein sequence is contained largely in one frame interrupted by three termination codons (Fig. 1). Dot matrix comparison of both protein sequences indicates a relatively high degree of similarity (Fig. 3), and the overall amino acid identity is about 46%. Comparisons of the overall S71 polymerase sequences with human endogenous provirus 4-1 (Repaske *et al.*, 1985) and with AKV (Etzerodt *et al.*, 1984) were also carried out (data not shown), confirming our previous results of sequence similarity between all three sequences (Leib–Mösch *et al.*, 1986). Computer analysis of various retroviral polymerase genes indicates that these are composed of five sections organized 5'-PR-RT-tether-RH-IN-3' (Johnson *et al.*, 1986). By comparison with polymerase genes from other retroviruses such as AKV, the SSAV/SSV pol sequence can be deduced to contain the tether-RH-IN sections (Fig. 4). The high degree of similarity between the SSAV/SSV and S71 amino acid sequences and collinearity of both sequences over more than 600 amino acids indicate that the S71 pol sequences are also organized 5'-tether-RH-IN-3'. The S71 pol sequence ends about 30 nucleotides upstream of the beginning of the envelope reading frame in SSV (Devare *et al.*, 1983).

The deduced amino acid sequence of the S71 tether section (Fig. 2 positions 3111–3489) is 126 amino acids in length and of these 66 (52%) are identical to



Fig. 4. Organization of retrovirus-related sequences in S71 as compared to full-length proviruses (AKV, SSAV) and a transforming defective provirus (SSV). The genomic organization of the murine endogenous AKV provirus was taken from the viral nucleotide sequence (Etzerodt *et al.*, 1984) and supplemented by comparison with the MoMuLV sequence (Shinnick *et al.*, 1981). Division of the AKV pol gene into PR, RT, tether, RH, and IN was adopted from Johnson *et al.* (1986). The organization of the truncated SSV provirus was determined from the published nucleotide sequence (Devare *et al.*, 1983). The relative location of SSV sequences in the SSAV genome was inferred from restriction sites common to both proviral sequences (Gelmann *et al.*, 1981; Devare *et al.*, 1983). The region of the SSAV pol gene sequenced for comparison with S71 is underlined in boldface. Where retroviral genes are either partially or completely missing (SSV and S71 pol, env) the remaining sequences are aligned with the corresponding AKV sequences and contiguous sequences are connected by slanted lines. S-NRF, S71-nonretroviral reading frame. B, *Bg/l*I; H, *Hind*III; K, *Kpn*I; P, *Pst*I; S, *Sst*I; Sa, *Sal*I; Xb, *Xba*I; Xh, *Xho*I. **■**, LTR; **■**, gag gene; **■**, polymerase gene; **■**, envelope gene; **□** in the SSAV provirus, unsequenced regions not contained in SSV.

the SSAV/SSV sequence. Comparisons of the amino acid sequences of various tether sections have suggested that this region is relatively poorly conserved between different retroviral polymerase genes (Johnson *et al.*, 1986) and is the fastest changing entity of retroviral polymerase genes (McClure *et al.*, 1988). However, we find the number of amino acid identities between the tether sections of AKV, BaEV, SSAV, and S71 to lie in the same range as the amino acid identities between the respective RH regions (Table 1) previously shown to be relatively highly conserved between various polymerase genes (Johnson *et al.*, 1986). This suggests that, within the group of C-type retroviruses, the

AMINO ACID IDENTITIES OF THE TETHER (T) AND RNASE H REGION OF pol (%)

	S [.]	S71		(A)V	Ва	EV	AKV		
	Т	RH	Т	RH	Т	RH	Т	RH	
S71			52	49	49	47	49	53	
SS(A)V				_	65	61	64	61	
BaEV					_		71	65	

tether region is as well conserved as other sections of the polymerase gene and may be of more significance than just a spacer separating the reverse transcriptase and RH domains.

By alignment with various other C-type retroviral RH sequences (see below) we found the S71 RH domain to extend from position 3489 to 3956 (Fig. 2). 82/156 amino acids in the S71 pol frame (Fig. 1) are identical with SSAV/SSV. Amino acid identity between both sequences is interrupted in a short stretch between position 3660 and 3707 in the S71 pol frame (frame 0). However, in this region the -1 reading frame (Fig. 1, frame 2) contains 11 contiguous amino acids identical with the SSAV/SSV pol sequence (Fig. 2). This increases the amino acid identities between both sequences to 60%. Levin et al. (1988) report the presence of conserved amino acids in the RH domain. All of these are also conserved in the S71 RH sequence (marked with an asterisk in Fig. 2) and three of these are located between position 3660 and 3707 in the -1 frame.

The sequences between position 3957 and 4938 constitute the IN section. Out of 21 invariant amino acid residues pointed out by Johnson *et al.* (1986), 13 are conserved in both the S71 and the SSAV/SSV se-



Fig. 5. Multiple alignment of a short conserved region within the endonuclease section of 13 retroviral polymerase sequences (GENALIGN): COPIA (aa 491–530), ERV3 (4980–5111), ERV1 (25–151), CHIMP (44–169), 4-1 (5365–5490), SSV (2936–3064), AKV (4996–5124), BaEV (4908–5036), S71 (4236–4364), MMTV (6006–6131), SMRV (202–328), RSV (4408–4533), HIV-1 (3999–4115). All numbers correspond to nucleotide positions, except COPIA (aa = amino acids). For references see Materials and Methods. Subgroup-specific conserved amino acids are shaded. The graph on the right shows amino acid identitics between consecutive sequence pairs (from top to bottom).

quence (marked by an asterisk in Fig. 2). The overall amino acid identity of the S71 IN sequence with the SSAV/SSV pol sequence is about 40%.

#### **Overall structure of S71**

Upon retroviral integration host cell sequences are duplicated so that the provirus is flanked by short direct repeats. A direct repeat of the 7-bp sequence immediately adjacent to the S71 LTR-like region is located 400 bp upstream of the gag-related region (positions 98-104 in Fig. 2). Assumption that this 7-bp duplication is a result of retroviral integration results in a length of 5406 nucleotides for the S71 element. Next to the gag (1478 nucleotides) and the pol regions (1818 nucleotides), the S71 element contains two regions with no discernible similarity to any retroviral sequences available for comparison. The first region spans the 394-bp between the 5' direct repeat and the gag-related region. The second retrovirus unrelated sequence is 1132 bp long and separates the S71 gag and pol regions (positions 1979-3110 in Fig. 2). Translation of this sequence suggests a possible open reading frame extending from a methionine codon at position 1979 to a termination codon at position 2344 (Fig. 2) and corresponding to 121 amino acids (13 kDa).

# Phylogenetic relationship of S71 to other retroviruses/retroviral elements

One approach to obtain information about phylogenetic relationships is to look for common mutations in a set of similar sequences. Composition of such sets is often limited by the number of available sequences. The IN section of retroviral polymerase genes seemed to be a likely candidate for this type of analysis since it has previously been shown to contain a number of highly conserved residues (Johnson *et al.*, 1986). Within the endonuclease section we chose a region comprising about 40 amino acids for which sequence data were available for 13 different retroviral elements. Multiple alignments were carried out using the GEN-ALIGN program and the results are shown in Fig. 5.

The SSV, AKV, BaEV, and S71 sequences all contain an Asp (D) residue at position 20 and Thr (T) at position 35 while ERV3, ERV1, CHIMP, and 4-1 exhibit a Cys (C) (ERV3 Tyr, Y) and a Lys (K) in the respective positions. This suggests that these C-type retroviral sequences can be divided into two subgroups. Careful analysis of the alignment reveals subgroup-specific amino acid preferences at 14 positions (2, 4, 6, 10, 13, 15, 20, 32, 33, 35, 38, 40, 41, and 43) at which identical amino acid residues are conserved in at least three of four members of a putative subgroup (shaded in Fig. 5).

S71 shares 7 out of the 8 amino acids specific for the SSV, AKV, BaEV subgroup, while it contains none of the conserved amino acids specific for the other subgroup. This suggests that S71 is more closely related to sequences AKV, BaEV, and SSV than to the other human endogenous retroviral elements ERV3, ERV1, and 4-1. pol

Chimp IN

ERV3 IN

IN

4-1

#### endonuclease region (IN) AKV IN IDFTEVKPGLYGYKYLLVFVDTFSGWVEAFPTKRETARVVSKK BaEV IN IDFTEVKPHYAGYKYLLVFVDTFSGWVEAFPTROETAHIVAKK SSV IN VDFTEVKPGRYGNRYLLVFDTFSGWVEAFPTRTETALTVCKR S71 IN ITE IKPHWAGYKYLLVLVDTFSGXTEAFATKNETATTVVKF ERVI IN GLYGI ASGGYRYMLVFVCTFSGWVGAFPTOTEKAGEVTOV

DFTKL PLVGGYRYMLVFVCTFSGWVEAFPTQTEKEQEVTQV

WTYRVAPGGSYRYMLVLVYTFSGWAKAFPTRSKNSXEVTK1

VDFTEM PKCGGNKYLLVLVCTYSGQVEAYPTRTEKAHEVTRV



## RNase H region (RH)

		Ŭ	*	•					* *	** *	<del>ب</del> ۲	**		
AKV	RH	PDADHTWY	TDGSSFI	Q EGORI	AGAA	VTTETEV	IWARA	LPAGI	SAOR	AELI/	LTQ.	ALKM	AEGKR	
BaEV	RH	PDADHTWY	TDGSSYL	D SGTRE	RAGAA	VVDGHNT	IWAQS	LPPGI	SAQK	AELI/	LTK	ALEL!	skgkk	
SSAV	RH	PGV PAW	TDGSSFI	A EGKRE	RAGAA	IDDGKRT	VWASS	LPEGI	SAQK	AELVA	LTQ	ALRL	AEGRD	
S71	RH	KEVDATVE	TDSSSLI	k ogvri	(AGAA)	VIMETOK	LQTQA	LPAG	SAOK	AELVA	<b>LIQ</b>	ALRR	VRTNV	
4-1	RH	SVDWELY	VDGSSFV	NPQEERO	CAGYA	VVTLDTV	AEARS	FPQG1	STOK	AELIA	<b>LIR</b>	ALEL:	SEGKT	
HIV	RH	IVGAETFY	VDGAAN	RETKI	LGKAG	YVTNKGR	QKVVP	L TNI	TNOK	TE LQI	ILL	alq i	DSGLE	
				•										
						*		* *	r	*			*	
AKV	RH	LINVYTDSF	RYAFATAH	IHGEIYE	RRGL	LTSEGRE	IKNKS	EILAI	LKAL	FLPKF	TS1	IHCL	Ghokgd	
BaEV	RH	ANIYTDSF	YAFATAH	THGSIYE	ERRGL	LTSEGKE	IKNKA	EIIAI	LKAL	FLPQE	IAVS	IHCPO	Ghokgo	
SSAV	RH	INIYTDSF	YAFATAH	IHGAIY	QRGLI	LTSAGKD	IKNKE	EILAI	LEAI	HLPKE	<b>WAI</b>	IHCPO	Ghqkgn	
S71	RH	LTFTLTAG	MLFATVF	VHGAIY	VRGLI	LTSAGKA	IKNXE	EILAI	LEAV	CLPQC	VAV	IHCK	Ghoked	
4-1	RH	VNIYTDS)	YVFLTLC	VHGALCI	(EKGL)	LNSGGKD	IKYQQ	EILQI	LEAV	KPHI	(VAV	IHCG	GHOXAS	
HIV	RH	VNIVTDS	YALGIIC	AQPD I	(SESE)	LVNQIIE	OLIKK	EKVYI	, A	WVPAH	IKGI	G	GNEQVD	
		*												
akv	RH	SAEARGNE	rladqaaf	EAAI	ľ	KTPPDTS	TLL							
BaEV	RH	DPVAVGNF	QADRVAF	QAAMAE \	/LTLA	reponts	HIT							
SSAV	RH	DPVATGNF	RADEAAK	QAAL	5	STRVLAE	TTK							
S71	RH	TAVAHGNO	RADSAAN	GPA	(	DLPVAPP	TLL							
4-1	RH	TLVGLGNS	CTDLEAC	KAAS	1	ALPGISD	SPP							
HIV	RH	KLVSAGIF	2		F	KILFLDG	IDK							



Fig. 6. Phylogenetic analysis of primate and murine retroviral sequences. Multiple alignments of amino acid sequences in two retroviral genes and corresponding phylogenetic trees shown on the right side: CA(s) (short): AKV (1296–1497), BaEV (1296–1457), ERV3 (1323–1484), SSV (2064–2225), S71 (1022–1282), 4-1 (1744–1908). CA(I) (long): BaEV (1266–1937), SSV (2034–2705), AKV (1306–1977), S71 (1092–1762), 4-1 (1717–2391), HIV-1 (807–1473). IN: BaEV (4908–5033), AKV (4996–5121), SSV (2936–3061), S71 (4236–4358), CHIMP (44–166), ERV1 (25–148), 4-1 (5365–5487), ERV3 (4980–5108). RH: AKV (4153–4623), BaEV (4077–4568), SSAV (614–1081), S71 (3489–3956), 4-1 (4534– 5004), HIV-1 (3429–3809). All numbers correspond to nucleotide positions. For references see Materials and Methods. In the alignments a dash indicates a frameshift, an X represents a stop codon. Amino acids that are conserved in all sequences are marked by an asterisk.

For additional confirmation, eight of the sequences most similar to S71 in Fig. 5 were used for phylogenetic analysis according to the method of Feng and Doolittle (1987). The corresponding phylogenetic trees clearly divide these sequences into the subgroups defined by common mutations in the GENALIGN alignment (Fig. 6, IN). However, the rather short range of the sequences compared could result in a biased alignment, as both methods use similar algorithms.

Another region known to be conserved is the RH domain of the polymerase (about 160 amino acids; Johnson *et al.*, 1986). The clustering order of AKV, BaEV, SSAV, S71, and 4-1 obtained with this much longer stretch of amino acids agrees with that of the IN sequences (Fig. 6, RH). Unfortunately, sequences of this region are not available for the other human endogenous elements ERV1 and ERV3. Furthermore, RH and IN are both part of the same gene known to be most highly conserved among retroviruses (Chiu *et al.*, 1984;

McClure et al., 1988). Therefore we decided to extend this analysis to sequences derived from an independent gene. For this purpose we used a conserved segment in CA for which sequence data were available for six of the eight elements used in the endonuclease alignment. The resulting phylogenetic tree (Fig. 6, CA(s)) shows essentially the same branching order as the two pol-derived trees (RH and IN). Furthermore, seguences 4-1 and ERV3 are clustered into a separate subgroup from S71, as in the endonuclease tree (IN). The relative order of the AKV, BaEV, SSV/SSAV, S71 and Chimp, ERV1, ERV3, 4-1 sequences, respectively, was found to be identical for all four multiple alignments. Extension of analysis to a much longer gag region (about 230 amino acids, including the highly conserved region) yielded essentially the same relative branching order as the short conserved CA region and the pol RH domain, the only difference being that SSV clustered with BaEV instead of AKV (Fig. 6). This sug-

#### gag

#### conserved CA region (short)

		* * *	**	*	*	*	*	**	**	***	*	*	*
AKV	CA(s)	LYNWKNN	NPSFSE	OPGE	(LTZ	LIESVLI	THOP	TWD	DCC	QLLG	TLLTG	EEK(	RVLL
BaEV	CA(s)	LYNWKTH	NPSFSQ	)PQ/	\LTS	LIESIL	THOP	TWD	DCC	OLLO	VLLTT	EER	RVLL
SSV	CA(s)	LYNWKSN	HPSFSEI	PAC	LTC	LLESLM	SHOP	TWD	DCC	OLLQ	ILFTT	EERI	ERILL
S71	CA(s)	LCN-KAH	NPSFSEI	(PQ)	/LTS	LMESVL	THOP	TWD	DCC	QLLL	TLFTS	EERI	<b>DRIRR</b>
4-1	CA(s)	LLNWKNN	TPSYTE	(PQ/	/LII	LLQTII	2THN₽	TWA	DCH	QLLM	FLFKT	PEXI	RVLQ
ERV3	CA(s)	LLNWKHHTP	-TPSYTEI	(PQ)	\LTI	LMOSIF	2TQN₽	TWP	DCF	QLLL	TLENT	EECI	RVTQ
EKA3	CA(S)	LLNWKHHTP-	-TPSYTEI	(PQ/	ALTI.	DIMOSTIN	21QNP	TWP	DC	QLLL	TLENT	EECI	KA10



conserved	CA region (	long)
-----------	-------------	-------

BAEV	CA(1)	IQYWPFSASDLYNWKTHNPSFSQDPQALTSLIESILLTHQPTWDDCQQLLQVLLTTEERQRVLL
SSV	CA(1)	LQYWPFSSADLYNWKSNHPSFSENPAGLTGLLESLMFSHOPTWDDCQQLLQILFTTEERERILL
AKV	CA(1)	LQYWPFSSSDLYNWKNNNPSFSEDPGKLTALIESVLTTHOPTWDDCQQLLGTLLTGEEKQRVLL
S71	CA(1)	LVYVPFSTSDLCN-KAHNPSFSEKPQVLTSLMESVLWTHOPTWDDCQQLLLTLFTSEERDRIRR
4-1	CA(1)	VYQPFTSADLLNWKNNTPSYTEKPQALIDLLQTIIQTHNPTWADCHQLLMFLFKTDERXRVLQ
HIV	CA(1)	VEEKAFSPEVIPMFSALSE GATPODLNTMLNTV GGHQAA MQMLKETINEEAAEWDRV
		*
BAEV	CA(1)	EARKNVPGPGGLP TQLPNEIDEGFPLTRPDWDYETAPGRESL RIYRQALLAGLKGAGKRPTNL
SSV	CA(1)	EARKNVLGDNGAP TQLENLINEAFPLNRPQWDYNTAAGRELL LVYRRTLVAGLKGAARRPTNL
AKV	CA(1)	EARKAVRGNDGRP TOLPNEVDAAFPLERPDWDYTTORGRNHL VLYROLLLAGLONAGRSPTNL
S71	CA(1)	EARKYFLTLAGRPEGEAQNLLEEVFPSTRPDXDPNSSGGKRAL DNFHRYLLAGIKGAARKPXNL
4-1	CA(1)	AATKWLEEHALADYONPOEYVRTQLPGTDPOWDPNXREDMORL NRYRKALLEGLKRRAOKATNI
HIV	CA(1)	HPVHACPIAPGQMREPRCSDIACTTSTLQEQICWMTNNPPIPVCEIYKRWIILCLNKIVRM YSP
		** * *
BAEV	CA(1)	** * AKVRTITQGKDESPAAFMERLLEGFRMYTPFDPEAPEHKATVAMSFIDQAALDIKGKLQRLDGIQ
BAEV SSV	CA(1) CA(1)	** * AKVRTITOGKDESPAAFMERLLEGFRMYTPFDPEAPEHKATVAMSFIDOAALDIKGKLORLDGIO AKVREVLQGPAEPPSVFLERLMEAYRRYTPFDPSEEGQQAAVATAFTGQSAPDIKKKLQRLEGLQ
BAEV SSV AKV	CA(1) CA(1) CA(1)	** * AKVRT I TOGKDE SPAAFMERLLEGFRMYTPFDPEAPEHKATVAMSF I DOAALD I KGKLORLDG I Q AKVREVLOGPAEPPSVFIERLMEAYRRYTPFDPSEEGOQAAVATAFTG SAPD I KKKLORLEGLO AKVKD I TOGPNE SPSAFLERLKEAYRRYTPYDPEDPGQETNVSMSF I WOSAPD I GRKLERLEDLK
BAEV SSV AKV S71	CA(1) CA(1) CA(1) CA(1)	** * AKVRTITOGKDESPAAFMERLLEGFRMYTPFDPEAPEHKATVAMSFIDQAALDIKGKLORLDGIQ AKVREVLQGPAEPPSVFLERLMEAYRRYTPFDPSEEGQQAAVATAFTGQSAPDIKKKLQRLEGLQ AKVKDITQGPNESPSAFLERLKEAYRRYTPYDPEDPCQETMVSMSFIMQSAPDIGRKLERLEDLK S-TTEVVQGPGEXPGAFLECLQEAYLTYTPFDPAAPEKSRVINLAFVAQ-ASDIRKKLQKLEGFA
BAEV SSV AKV S71 4-1	CA(1) CA(1) CA(1) CA(1) CA(1) CA(1)	** * AKVRTITOCKDESPAAFMERLLEGFRMYTPFDPEAPEHKATVAMSFIDOAALDIKGKLORLDGIO AKVREVLQGPAEPPSVFLERLMEAYRRYTPFDPSEEGQQAAVATAFTGQSAPDIKKKLQRLEGLQ AKVKDITOGPNESPSAFLERLKEAYRRYTPYDPEDPCQETNVSMSFINQSAPDIGRKLERLEDLK S-TTEVVQGPGEXPGAFLECLQEAYLTYTPFDPAAPEKSRVINLAFVAQ-ASDIRRKLQKLEGFA NKVSEVIQGKEESPAKFHERLCEAYCMYTPFDPDSPENQMINMALVSQSTEDIRRKLQKKAGFA
BAEV SSV AKV S71 4-1 HIV	CA(1) CA(1) CA(1) CA(1) CA(1) CA(1)	** * AKVRTITOGKDESPAAFMERLLEGFRMYTPFDPEAPEHKATVAMSFIDOALDIKGKLORLDGIO AKVREVLOGPAEPPSVFLERLMEAYRRYTPFDPSEEGOQAAVATAFTGOSAPDIKKKLORLEGLO AKVKDITOGPNESPSAFLERLKEAYRRYTPYDPEDPGOETNVSMSFIMOSAPDIGRKLERLEDLK S-TTEVVQGPGEXPGAFLECLOEAYLTYTPFDPAAPEKSRVINLAFVAQ-ASDIKKLOKLEGFA NKVSEVIQGKEESPAKFHERLCEAYCMYTPFDPDSPENORMINMALVSGSTEDIRKKLOKKAGFA TSILDIRQGPKEPFRDYVDRF YKTLRAEQASQEVKNMMTETLLVQNANPDCKTILKALG P
BAEV SSV AKV S71 4-1 HIV	CA(1) CA(1) CA(1) CA(1) CA(1) CA(1)	** * AKVRT I TOGKDE SPAAFMER LLEGFRMYTPF DPE APEHKATVAMSF I DOAALD I KGKLORLDG I O AKVREVLOGPAEPPS VFLERLMEAYRRYTPF DPSEEGOO AAVATAF TOG SAPD I KKKLORLEGLO AKVKD I TOG PNE SPSAF LERLKEAYRRYTP YDPE DPG QE TNVSMSF I WOSAPD I GRKLERLED LK S-TTEVVOG PGEXPGAF LECLOEAYLTYTPF DPAAPEKSRVINLAFVAO-ASD I RKKLOKLEGFA NKVSEVI QG KEESPAKFHERLCEAYCMYTPF DPD SPENORMI NMALVSQSTED I RRKLOKKAGFA TSILD I ROG PKEPFRD YVDRF YKTLRAE QAS QEVKNWMTET LLVONANPDCKTI LKALG P
BAEV SSV AKV S71 4-1 HIV	CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1)	** * AKVRTITOGKDESPAAFMERLLEGFRMYTPFDPEAPEHKATVAMSFIDOAALDIKGKLORLDGIQ AKVREVLOGPAEPPSVFIERLMEAYRRYTPFDPSEEGQQAAVATAFTGQSAPDIKKKLORLEGLQ AKVKDITOGPNESPSAFLERLKEAYRRYTPYDPEDPGQETNVSMSFIMQSAPDIGKKLERLEDLK S-TTEVVOGFGEXFGAFIECLOEAVLTYTPFDPAAPERSKVINLAFVAQ-ASDIRKKLOKLEGFA NKVSEVIQGKEESPAKFHERLCEAYCMYTPFDPDSPENQRMINMALVSQSTEDIRRKLOKKAGFA TSILDIRQGPKEPFRDYVDRF YKTLRAEQASQEVKNMMTETLLVQNANPDCKTILKALG P
BAEV SSV AKV S71 4-1 HIV BAEV	CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1)	** * AKVRTITOGKDESPAAFMERLLEGFRMYTPFDPEAPEHKATVAMSFIDOAALDIKGKLORLDGIO AKVREVLOGPAEPPSVFIERLMEAYRRYTPFDPSEEGOOAAVATAFTGOSAPDIKKKLORLEGLO AKVKDITOGPNESPSAFIERLKEAYRRYTPYDPEDPCOETNVSMSFINGSAPDIGRKLERLEDLK S-TTEVVQGPGESPSAFIERLCEAYCMYTPFDPAAPEKSRVINLAFVAQ-ASDIRKKLOKLERLEDLK NKVSEVIQGKEESPAKFHERLCEAYCMYTPFDPDSPENORMINMALVSQSTEDIRRKLOKKAGFA TSILDIRQGPKEPFRDYVDRF YKTLRAEQASQEVKNMMTETLLVQNANPDCKTILKALG P THCLQELVREAE KVYNKRETPEEREARLIKEQ
BAEV SSV AKV S71 4-1 HIV BAEV SSV	CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1)	** * AKVRTITOGKDESPAAFMERLLEGFRMYTPFDPEAPEHKATVAMSFIDOAALDIKGKLORLDGIO AKVREVLOGPAEPPSVFLERLMEAYRRYTPFDPSEEGOOAAVATAFTGOSAPDIKKKLORLEGLO AKVKDITOGPNESPSAFLERLKEAYRRYTPYDPEDPGOETNVSMSFIMOSAPDIGRKLERLEDLK S-TTEVVOGPGEXPGAFLECLOEAYLTYTPFDPAAPEKSRVINLAFVAQ-ASDIKKLOKLEGFA NKVSEVIOGKEESPSAKFHERLCEAYCMYTPFDPSPENORMINMALVSGSTEDIRKKLOKLEGFA TSILDIROGPKEPFRDYVDRF YKTLRAEQASQEVKNMMTETLLVONANPDCKTILKALG P THGLQELVREAE KVYNKRETPEEREARLIKEQ DYSLQDLVREAE KVYNKRETPEEREARLIKEQ
BAEV SSV AKV S71 4-1 HIV BAEV SSV AKV	CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1)	** * AKVRTITOGKDESPAAFMERLLEGFRMYTPFDPEAPEHKATVAMSFIDOAALDIKGKLORLDGIQ AKVREVLOGPAEPPSVFLERLMEAYRRYTPFDPSEEGQQAAVATAFTGQSAPDIKKKLORLEGLO AKVKDITOGPNESPSAFLECLQEAYLTYTPFDPAPESRVINLAFVQ-ASDIRKKLORLEGLO AKVKDITOGPNESPSAFLECLQEAYLTYTPFDPAPERSVINLAFVQ-ASDIRKKLORLEGLO NKVSEVIQGKEESPAKFHERLCEAYCMYTPFDPAPERSVINLAFVQ-ASDIRKKLOKLEGFA NKVSEVIQGKEESPAKFHERLCEAYCMYTPFDPDSPENORMINMALVSQSTEDIRRKLOKKAGFA TSILDIRQGPKEPFRDYVDRF YKTLRAEQASQEVKNMMTETLLVQNANPDCKTILKALG P THGLQELVREAE KVYNKRETPEEREARLIKEQ DYSLQDLVREAE KVYNKRETPEEREARLIKEQ SKTLGDLVREAE RIFNKRETPEEREARLIKEQ
BAEV SSV AKV S71 4-1 HIV BAEV SSV AKV S71	CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1)	** * AKVRTITOGKDESPAAFMERLLEGFRMYTPFDPEAPEHKATVAMSFIDOAALDIKGKLORLDGIO AKVREVLOGPAEPPSVFIERLMEAYRRYTPFDPSEEGQQAAVATAFTGQSAPDIKKKLORLEGLQ AKVKDITOGPNESPSAFIERLKEAYRRYTPYDPEDPOQETNVSMSFINGSAPDIGKLERLEDLK S-TTEVVQGPGEXPGAFIECLOEAYLTYTPFDPAAPEKSRVINLAFVAQ-ASDIRKKLOKLEGLA NKVSEVIQGKEESPAKFHERLCEAYCMYTPFDPDSPENORMINMALVSQSTEDIRRKLOKKAGFA TSILDIRQGPKEPFRDYVDRF YKTLRAEQASQEVKNMMTETLLVQNANPDCKTILKALG P THGLQELVREAE KVYNKRETPEEREARLIKEQ DYSLQDLVREAE RIFNRETFEERERERKKEA SKTLGDLVREAE RIFNRETFEERERERKRET GMNISQLLEVAQ RIFDSQEFEKQKQAAEKAAD
BAEV SSV AKV S71 4-1 HIV BAEV SSV AKV S71 4-1	CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1) CA(1)	** * AKVRTITOGKDESPAAFMERLLEGFRMYTPFDPEAPEHKATVAMSFIDOAALDIKGKLORLDGIO AKVREVLQGPAEPPSVFLERLMEAYRRYTPFDPSEEGQQAAVATAFTGQSAPDIKKKLQRLEGLQ AKVKDITOGPNESPSAFIERLKEAYRRYTPYDPEDPCQETNVSMSFIMOSAPDIGKKLERLEDLK S-TTEVVQGPGEXPGAFLECLQEAYLTYTPFDPAAPEKSRVINLAFVAQ-ASDIKKLQKLERGFA NKVSEVIQGKEESPAKFHERLCEAYCMYTPFDPDSPENQMINMALVSQSTEDIRKKLQKKAGFA TSILDIRQGPKEPFRDYVDRF YKTLRAEQASQEVKNMMTETLLVQNANPDCKTILKALG P THGLQELVREAE KVYNKRETPEEREARLIKEQ DYSLQDLVREAE KVYNKRETPEEREERVREA SKTLGDLVREAE RIFNKRETPEEREERVREA GNNISQLLEVAQ RIFDSQEFEKQRQAAEKAAD GNNISQLLEVAQ RIFDSQEFEKQRQAAEKAAD



gests that, at least in our case, these short ranges of amino acid sequences (IN, CA) reflect the overall relationship obtained with three to four times larger regions (RH, CA).

Each tree shown in Fig. 6 is based on calculations from the distance matrix derived from binary alignments. While the overall topology remains unaltered over a considerable range of distances, the individual branch lengths are highly dependent on the assumption of common rates of amino acid changes for the sequences compared. Since defective endogenous sequences do not face continuous functional selection, a common rate of amino acid changes with active viral sequences cannot be assumed. The branch lengths in Fig. 6 are plotted according to the calculated distances in the depicted trees without further adjustments. Although this may not correlate with the real evolutionary distances, it should not interfere with the overall topology of the trees.

#### DISCUSSION

The organization of retrovirus-related sequences in the S71 element was determined by nucleotide se-

quence analysis and comparison with the corresponding SSV and SSAV/SSV sequences. The S71 truncated proviral element contains gag- and pol-related sequences followed by a 3' LTR-like sequence. The S71 gag region is of roughly the same length as the SSV gag gene (1478 vs 1539 nucleotides) and shows distinct sequence similarity to all sections of the SSV gag gene except p12gag. A similar situation can also be observed for the gag region of the human endogenous Ctype provirus 4-1 which also exhibits least homology to MoMuLV gag within the p12gag section (Repaske et al., 1985). On the basis of the alignment of the amino terminal CA sequences, Oroszlan et al. (1977, 1981) divided mammalian C-type retroviruses into four subgroups. Comparison of the above region in S71 with the sequences characteristic for each of the four subgroups shows that the S71 element clearly falls into the SSAV/GaLV subgroup (Fig. 7).

CA(I)

The S71 pol region consists of tether–RH- and INrelated sections. The retrovirus-related sections within the gag and pol regions of S71 are arranged in the relative order typical of a number of murine and primate retroviral genomes. However, the S71 element is lack-

#### WERNER ET AL.



Fig. 7. Comparison of CA amino terminal amino acids according to Oroszlan *et al.* (1977, 1981) in Weiss *et al.* (1985). Ra-MLV, Rauscher murine leukemia virus; Mo-MLV, Moloney murine leukemia virus; FeLV, feline leukemia virus; RD-114, feline endogenous virus; GaLV, Gibbon ape leukemia virus; MMC-1, endogenous type C virus of *Macaca mulatta;* MAC-1, endogenous type C virus of *M. arcotoides,* CPC-1, endogenous type C virus of *Colobus polytomos.* X, unidentified amino acid.

ing over half of the polymerase gene, encompassing the protease coding region and the complete reverse transcriptase domain (Fig. 4).

A prevalent feature of transducing retroviruses is the partial or complete lack of genes required for retroviral multiplication, as is the case in S71. The sequence relatedness of S71 with the v-sis-bearing defective SSV provirus in the gag and endonuclease regions suggests comparison of the overall organization of retroviral sequences in both proviral elements. Both elements contain a complete gag gene, whereas both are missing a large part of the polymerase gene, including the entire reverse transcriptase domain, and most or all of the envelope gene (Fig. 4). Notably, in both elements part of a retroviral gene has been replaced by a presumably nonretroviral sequence. In SSV part of the envelope gene is replaced by the sis oncogene which is essential for the SSV transforming activity (Robbins et al., 1982). Transcription of v-sis can occur in the absence of the 5' LTR and donor splice signals suggesting that appropriate signals reside in the SSAV sequences preceding v-sis (Devare et al., 1983). As mentioned above, part of the pol gene in S71 is replaced by a presumably nonretroviral sequence with a 121 amino acid open reading frame (S71-nonretroviral reading frame, S-NRF; Fig. 4). We have recently obtained evidence for the expression of related sequences in human placenta and are currently investigating the possibility that S-NRF may have a cellular counterpart not associated with retrovirus-related sequences (manuscript in preparation). Therefore, similarities between S71 and SSV are not confined to sequence level but can be expanded to include their general structure implicating analogous mechanisms for the generation of both truncated proviral sequences.

Direct repeats flanking integrated proviruses are known to be a hallmark of the retroviral integration mechanism. Although the exact number of duplicated nucleotides cannot be determined for S71, their presence and location in S71 suggests that S71, like SSV, is a proviral element resulting from a retroviral integration event. In this case the absence of cis sequences in S71 required for integration makes it conceivable that at some point there existed a full-length counterpart to S71 analogous to SSAV/SSV.

Sequence comparison of S71 with SSV/SSAV by dot matrix analysis indicates a higher degree of conservation on amino acid level than on nucleotide level (Fig. 3). This was analyzed in more detail for the short conserved CA and IN regions (Table 2). Only one amino acid substitution in the gag region and none in the endonuclease region is caused solely by a nucleotide change in the third position. The distribution of nucleotide substitutions among the three positions of the respective codons therefore are clearly biased towards silent third position changes. This bias suggests func-

TA	BL	E	2
----	----	---	---

#### NUCLEOTIDE SUBSTITUTIONS VERSUS AMINO ACID CHANGES IN S71 AS COMPARED TO SSV

R	egion:	1. Position	2. Position	3. Position
CA IN		16 (25%) 14 (29%)	11 (18%) 9 (18%)	35 (56%) 27 (55%)
	Total:	Nucleotide substitutions	Changed codons	Amino acid substitutions
CA		62	37	19
IN		51	32	16

tional constraints effective on protein level indicating transient functionality of these proteins.

A critical factor in phylogenetic analysis with short range sequences is the choice of sequence regions used for comparison. We have identified two regions leading to results that reflect the overall situation as correctly as alignments with much longer sequences at least for C-type retrovirus-related sequences. Since our analysis demonstrates the reliability of these short sequence ranges they can be used as targets for polymerase chain reactions (PCR) for rapid classification of newly isolated sequences (Shih *et al.*, 1989). The initial phylogenetic evaluation of new elements could be carried out with these amplified sequences suggesting that the effort of large scale sequencing may no longer be necessary for this purpose.

Extension of this type of phylogenetic analysis to include sequences derived from the envelope gene would be a logical continuation. However, phylogenetic trees derived from env-related sequences differ widely from the largely congruent trees obtained with retroviral enzyme sequences such as RH and IN (Mc-Clure et al., 1988) as well as with the capsid proteinderived sequences analyzed in this study (Fig. 6). This suggests that, during retrovirus evolution, these three proteins have faced similar pressure for functional selection which is different from that governing evolution of envelope proteins (Doolittle et al., 1989). The rapid evolutionary change of envelope-related proteins is apparently at least in part caused by major genetic recombination events (McClure et al., 1988). Furthermore, exposure of envelope proteins to the immune system of the host calls for adaptive changes in these proteins, as reflected by the high rates of change in the envelope sequences of different HIV isolates.

We previously reported that the control regions of the S71 3'-LTR-like sequence are most homologous to infectious murine and primate proviruses (Brack-Werner et al., 1989a). All phylogenetic trees obtained with highly conserved retroid sequences within S71 also point out that S71 is more closely related to murine and primate infectious proviruses than to ERV1, ERV3, Chimp, and 4-1. Moreover the endonuclease (IN) and the short CA region (CA(s)) alignments show an even more distinct pattern clustering S71 with AKV, BaEV, and SSV into one subgroup and the other three (two) human endogenous retroviral sequences into another subgroup (Fig. 6). The relative order of all human endogenous elements compared to the murine and primate infectious proviruses is reproduced in each of the four alignments. We have reported the SSAV-related subgroup to comprise about 35 copies per human haploid genome (Brack-Werner et al., 1989b). Low stringency hybridization of human DNA with a specific S71 probe derived from the tether-RH region revealed at least 15 strongly hybridizing fragments (data not shown). On the basis of these findings, S71 may represent a subgroup of human endogenous retroviral elements which is distinct from the ERV1, ERV3, 4-1 subgroup. Sequencing of further endogenous retroviral elements is required to shed more light on the evolutionary relationship of human retroposons.

#### ACKNOWLEDGMENTS

We are grateful to R. F. Doolittle for his cooperation especially in providing us with his computer program package for multiple sequence alignments and construction of phylogenetic trees. It is a pleasure to thank Marion Ohlmann and Richard Albang for expert technical assistance. This work was supported by the Deutsche Forschungsgemeinschaft (Grant SFB 324) and by Contract BI6-156D of the Commission of the European Communities.

#### REFERENCES

- BONNER, T. I., O'CONNELL, C., and COHEN, M. (1982). Cloned endogenous retroviral sequences from human DNA. *Proc. Natl. Acad. Sci.* USA 79, 4709–4713.
- BRACK-WERNER, R., BARTON, D. E., WERNER, T., FOELLMER, B. E., LEIB-MÖSCH, C., FRANCKE, U., ERFLE, V., and HEHLMANN, R. (1989a). Human SSAV-related endogenous retroviral element: LTR-like sequence and chromosomal localization to 18q21. *Genomics* 4, 68– 75.
- BRACK-WERNER, R., LEIB-MÖSCH, C., WERNER, T., ERFLE, V., and HEHLMANN, R. (1989b). Human endogenous retrovirus-like sequences. *In* "Modern Trends in Human Leukemia VIII" (R. Neth, Ed.), Springer-Verlag, Berlin/Heidelberg/New York, in press.
- BRACK-WERNER, R., WERNER, T., LEIB-MÖSCH, C., HEHLMANN, R., and ERFLE, V. (1989c). Primary structure of the SSAV tether—RNase H—endonuclease (pol) region deleted in SSV. *Nucleic Acids Res.* **17**, in press.
- CALLAHAN, R. (1988). Two families of human endogenous retroviral genomes. *Banbury Rep.* **30**, 91–100.
- CHEN, E. Y., and SEEBURG, P. H. (1985). Supercoil sequencing: A fast and simple method for sequencing plasmid DNA. *DNA* **4**, 165– 170.
- CHIU, I.-M., CALLAHAN, R., TRONICK, S. R., SCHLOM, J., and AARONSON, S. A. (1984). Major pol gene progenitors in the evolution of oncoviruses. *Science* **223**, 364–370.
- COVEY, S. N. (1986). Amino acid sequence homology in gag region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. *Nucleic Acids Res.* **14**, 623–633.
- DEVARE, S. G., REDDY, E. P., LAW, J. D., ROBBINS, K. C., and AARONSON, S. A. (1983). Nucleotide sequence of the simian sarcoma virus genome: Demonstration that its acquired cellular sequences encode the transforming gene product p28(sis). *Proc. Natl. Acad. Sci. USA* 80, 731–735.
- DOOLITTLE, R. F., FENG, D.-F., JOHNSON, M. S., and MCCLURE, M. A. (1989). Origins and evolutionary relationships of retroviruses. *Q. Rev. Biol.* 64, 1–30.
- ETZERODT, M., MIKKELSEN, T., PEDERSEN, F. S., KIELDGAARD, N. O., and JØRGENSEN, P. (1984). The nucleotide sequence of the AKV murine leukemia virus genome. *Virology* **134**, 196–207.
- FENG, D.-F., and DOULTTLE, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25, 351–360.
- FÜTTERER, J., and HOHN, T. (1987). Involvement of nucleocapsids in reverse transcription: A general phenomenon? *Trends Biochem. Sci.* **12**, 92–95.

- GELMANN, E. P., WONG–STAAL, F., KRAMER, R. A., and GALLO, R. C. (1981). Molecular cloning and comparative analyses of the genomes of simian sarcoma virus and its associated helper virus. *Proc. Natl. Acad. Sci. USA* 78, 3373–3377.
- HANSEN, J., SCHULZE, T., MELLERT, W., and MÖLLING, K. (1988). Identification and characterization of HIV-specific RNase H by monoclonal antibody. *EMBO J.* **7**, 239–243.
- HIZI, A., and HUGHES, S. H. (1988). Expression of Moloney murine leukemia virus and human immunodeficiency virus integration proteins in *Escherichia coli. Virology* 167, 634–638.
- JOHNSON, M. S., MCCLURE, M. A., FENG, D.-F., GRAY, J., and DOOLIT-TLE, R. F. (1986). Computer analysis of retroviral pol genes: Assignment of enzymatic functions to specific sequences and homologies with nonviral enzymes. *Proc. Natl. Acad. Sci. USA* 83, 7648– 7652.
- KATO, S., MATSUO, K., NISHIMURA, N., TAKAHASHI, N., and TAKANO, T. (1987). The entire nucleotide sequence of baboon endogenous virus DNA: A chimeric genome structure of murine type C and simian type D retroviruses. *Japan. J. Genet.* 62, 127–137.
- KOZAK, M. (1984). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acid Res.* 12, 857–872.
- LEIB-MÖSCH, C., BRACK, R., WERNER, T., ERFLE, V., and HEHLMANN, R. (1986). Isolation of an SSAV-related endogenous sequence from human DNA. *Virology* **155**, 666–677.
- LEIS, J., BALTIMORE, D., BISHOP, M., COFFIN, J., FLEISSNER, E., GOFF, S. P., OROZLAN, S., ROBINSON, H., SKALKA, A. M., TEMIN, H. M., and VOGT, V. (1988). Standardized and simplified nomenclature for proteins common to all retroviruses. *J. Virol.* 62, 1808–1809.
- LEVIN, J. G., CROUCH, R. J., POST, C., HU, S. C., MCKELVIN, D., ZWEIG, M., COURT, D. L., and GERWIN, B. I. (1988). Functional organization of the murine leukemia virus reverse transcriptase: Characterization of a bacterially expressed AKR DNA polymerase deficient in RNase H activity. J. Virol. 62, 4376–4380.
- MCCLURE, M. A., JOHNSON, M. S., FENG, D.-F., and DOOLITTLE, R. F. (1988). Sequence comparison of retroviral proteins: Relative rates of changes and general phylogeny. *Proc. Natl. Acad. Sci. USA* 85, 2469–2473.
- MOORE, R., DIXON, M., SMITH, R., PETERS, G., and DICKSON, C. (1987). Complete nucleotide sequence of the milk transmitted mouse mammary tumor virus: Two frameshift suppression events are required for the translation of gag and pol. J. Virol. 61, 480–490.
- MOUNT, S. M., and RUBIN, G. M. (1985). Complete nucleotide sequence of the drosophila transposable element Copia: Homology between Copia and retroviral proteins. *Mol. Cell. Biol.* 5, 1630– 1638.
- O'CONNELL, C., O'BRIEN, S., NASH, W. G., and COHEN, M. (1984). ERV3, a full-length human endogenous provirus: Chromosomal localization and evolutionary relationship. *Virology* **138**, 225–235.
- OROSZLAN, S., COPELAND, T., SMYTHERS, G., SUMMERS, M. R., and GIL-DEN, R. V. (1977). Comparative primary structure analysis of the

p30 protein of woolly monkey and gibbon type C viruses. *Virology* **77**, 413–417.

- OROSZLAN, S., COPELAND, T. D., GILDEN, R. V., and TODARO, G. J. (1981). Structural homology of the major internal proteins of endogenous type C viruses of two distantly related species of old world monkeys: *Macaca arcoides and Colobus polykomos*. *Virology* **115**, 262–271.
- RATNER, L., HASELTINE, W., PATARCA, R., LIVAK, K. J., STARCICH, B., JOSEPHS, S. F., DORAN, E. R., RAFALSKI, J. A., WHITEHORN, E. A., BAUMEISTER, K., IVANOFF, L., PETTEWAY, JR., S. R., PEARSON, M. L., LAUTENBERGER, J. A., PAPAS, T. S., GHRAYEB, J., CHANG, N. T., GALLO, R. C., and WONG–STAAL, F. (1985). Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature (London) **313**, 277–284.
- REPASKE, R., STEELE, P. E., O'NEILL, R. R., RABSON, A. B., and MARTIN, M. A. (1985). Nucleotide sequence of a full-length human endogenous retroviral segment. J. Virol. 54, 764–772.
- ROBBINS, K. C., DEVARE, S. G., REDDY, E. P., and AARONSON, S. A. (1982). *In vivo* identification of the transforming gene product of simian sarcoma virus. *Science* 218, 1131–1133.
- SCHWARTZ, D. E., TIZARD, R., and GILBERT, W. (1983). Nucleotide sequence of Rous sarcoma virus. *Cell* 32, 853–869.
- SHIH, A., MISRA, R., and RUSH, M. G. (1989). Detection of multiple, novel reverse transcriptase coding sequences in human nucleic acids: Relation to primate retroviruses. J. Virol. 63, 64–75.
- SHINNICK, T. M., LERNER, R. A., and SUTCLIFFE, G. J. (1981). Nucleotide sequence of Moloney murine leukemia virus. *Nature (London)* 293, 543–548.
- SKOWRONSKI, J., and SINGER, M. F. (1986). The abundant LINE-1 family of repeated DNA sequences in mammals. *Cold Spring Harbor Symp. Quant. Biol.* 2, I, 457–464.
- TABOR, S., and RICHARDSON, C. C. (1987). DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc. Natl. Acad. Sci. USA* **84**, 4767–4772.
- TEICH, N. (1984). Taxonomy of retroviruses. *In* "Molecular Biology of Tumor Viruses, RNA Tumor Viruses" (R. Weiss, N. Teich, H. Varmus, and J. Coffin, Eds.), 2nd Ed., pp. 25–207, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
- TEMIN, H. M. (1985). Reverse transcription in the eukaryotic genome: Retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Mol. Biol. Evol.* 2, 455–468.
- VIEIRA, J., and MESSING, J. (1982). The pUC plasmids, an M13mp7derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* 19, 259–268.
- WEINER, A. M., DEININGER, P. L., and EFSTRATIADIS, A. (1986). Nonviral retroposons: Genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* 55, 631–661.
- WEISS, R., TEICH, N., VARMUS, H., and COFFIN, J. (Eds.) (1985). Appendix E: Amino acid sequences of retroviral structural proteins. *In* "Molecular Biology of Tumor Viruses, RNA Tumor Viruses," 2nd ed., Supplements and Appendixes, pp. 1215–1216, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.