The rat leukemia virus (RaLV) is an endogenous retrovirus that is spontaneously released by Sprague- Dawley rat embryo cells. The overall structure of the RaLV genome resembles that of other simple, replication- competent retroviruses, but the sequence of the long terminal repeats (LTR) is unique and unrelated to the known retroviruses. Phylogenetically, the RaLV genome appears to be more closely related to the feline leukemia virus group of retroviruses than to the murine leukemia virus group. A remarkable feature of RaLV is that it is capable of transducing a ras proto-oncogene from rat tumor cells in the form of an acutely transforming virus, designated the Rasheed strain of the rat sarcoma virus (RaSV). With the exception of the c-ras sequence, the genomes of both RaLV and RaSV are collinear. The RaSV-encoded oncogene v-Ra-ras expresses a fusion protein with a molecular mass of 29 kDa, and it exhibits a unique structure that has not been described previously for any known virus. The 3' end of this gene is derived from sequences encoding RaLV matrix followed by 20 bp derived from the U5 region of the RaLV LTR (RS-U5 element) which is joined at its 3' end to sequences derived from all six (coding and noncoding) exons of the c-ras proto-oncogene at the 3' end. This recombinational event represents a novel mechanism among the acutely transforming viruses that have been studied.

The rat leukemia virus (RaLV) is a mammalian type C retrovirus that is spontaneously released by a Sprague-Dawley (SD-1) rat embryo culture in vitro (41). Sequences related to RaLV are present in the normal rat genome but not in the DNAs of other mammals (including mice) or birds, indicating that RaLV is an endogenous retrovirus of the rat (41). This virus replicates well in cells of its original host (SD-1 cells) and those derived from different strains of rats but not in the cells of other vertebrates or humans; i.e., it is ecotropic (41).

Maintenance of RaLV-producing SD-1 cells or exogenously RaLV-infected normal rat embryo fibroblasts derived from other rat strains in vitro for 7 to 9 months (i.e., 40 to 50 passages) induces morphologic changes in chronically infected cells (41). This transformation process can be accelerated by treatment of RaLV-infected cells with chemicals or inhibited by growth of these cells in the presence of RaLV antibodies (42, 44). The RaLV produced by transformed SD-1 rat cells (SD-1T cells) is not acutely transforming to other rat embryo cells in vitro (44). A unique property of RaLV that distinguishes it from most mammalian retroviruses is its ability to transduce in vitro an activated c-ras gene to form an acutely transforming virus, designated the Rasheed rat sarcoma virus (RaSV), during cocultivation of SD-1T cells with cells isolated from various chemically induced tumors of rats (44). No acutely transforming virus was recovered when SD-1T cells were cocultivated with non-virus-producing cell lines derived from mouse, cat, primate, or human tumors (29, 44). Moreover, neither ras nor any other oncogene can be rescued when any of the same rat tumor cells which gave rise to the RaSV isolates are cocultivated with (i) rat or mouse cells infected by either ecotropic or amphotropic murine leukemia virus (MuLV) strains (43), (ii) human cells infected with either amphotropic MuLV or the endogenous cat virus RD114 (45), or (iii) cat cells infected with feline leukemia virus (FeLV) (43, 45). These data indicate that RaLV released by SD-1T cells plays a significant role in the transduction of ras oncogenes from rat tumors and in the biogenesis of RaSV in vitro (40).

The highly oncogenic or acutely transforming viruses have been recovered from naturally occurring solid tumors of various vertebrates or they have been generated in tumors produced in susceptible laboratory animals by repeated in vivo inoculations of leukemogenic viruses or lymphoma tumor extracts containing these viruses in rats or mice. For instance, in vivo inoculation of different rat strains with either Moloney murine leukemia virus (Mo-MuLV) or murine erythroleukemia virus and repeated in vivo passages of the lymphomas induced by these viruses in rats have resulted in the isolation of the Harvey murine sarcoma virus (Ha-MSV) (18) and Kirsten murine sarcoma virus (Ki-MSV) (26), respectively. Biogenesis of RaSV in vitro therefore represents an extraordinarily rare phenomenon worthy of further studies.

Both RaSV-transformed cells and RaSV-induced fibrosarcomas in rats express a transforming protein of 29 kDa which has been designated p29 or v-Ra-Ras (56, 57). An important characteristic of RaSV-induced transformation is that cells transformed with different isolates of RaSV derived from different rat tumors express the same or immunologically similar p29 proteins (57). Our earlier studies have suggested that although the RaSV p29 protein is related to the v-Ha-ras-encoded 21-kDa protein (p21), the biologic, immunologic, and
molecular properties of p29 are distinct from those displayed by the p21 protein that is expressed in normal or Ha-MSV-transformed cells (56, 57). Whereas the p29 oncoprotein can be immunoprecipitated from RaSV-transformed cells by both RaLV Gag- and p21-reactive antibodies, the transforming protein of Ha-MSV can be precipitated only by p21-specific antibodies (56, 57). In addition, cells transformed by Ha-MSV do not express p29 and RaSV-transformed cells do not express p21 (57). These studies have indicated that the RaSV-induced oncoprotein is a fusion product of the proteins encoded by the RaLV gag gene and the ras cellular proto-oncogene (40, 57).

The precise mechanisms involved in the generation of highly oncogenic viruses are not well understood. However, based on nucleotide sequence comparisons of the genomes of Ha-MSV, Ki-MSV, and the malignant histocytosis sarcoma virus of the mouse, it has been proposed that virus-like 30S RNA (VL30) elements that flank the 5’ and 3’ ends of normal c-ras proto-oncogenes of both the rat and mouse genomes are responsible for the recombination and transduction of ras oncogenes by the respective helper MuLV strains associated with these sarcoma viruses (14, 32). The presence of VL30 sequences in Ha-MSV, Ki-MSV, and malignant histocytosis sarcoma virus of the mouse has also been considered necessary for the expression of transforming activities by these viruses (14).

To identify the viral and cellular sequences that may be involved in the transduction of the rat c-ras proto-oncogene by RaLV, we sequenced an unintegrated RaLV DNA intermediate as well as a proviral RaSV genome (16). Comparisons of RaSV and RaLV sequences to each other and with those of MuLV, Ha-MSV, Ki-MSV, and c-ras proto-oncogenes that had been cloned from genomic DNAs of several normal rat cell cultures indicate several remarkable properties of both RaLV and RaSV genomes that have not been reported previously for other oncogenic or nononcogenic retroviruses.

Herein, we present the first report of the complete RaLV genome structure as well as the sequence of the RaSV proviral DNA. Our studies demonstrate that VL30-like sequences are not involved in the generation of RaSV and that homologous pairing of unique nucleotide sequences from two different regions of the RaLV genome are responsible for the recombination with all six of the coding as well as noncoding exons of the c-ras proto-oncogene to form RaSV.

This work was conducted by S.-Y. Lee in partial fulfillment of the requirements for a Ph.D. from the University of Southern California, Los Angeles, (1999).

MATERIALS AND METHODS

RaLV and RaSV plasmids. The details of the molecular cloning of the RaLV and RaSV genomes into pBR322 have been described previously (16). Briefly, the cloned RaLV genome was derived from an unintegrated, circular RaLV DNA intermediate contained within a Hirt supernatant of normal rat kidney (NRK) cells shortly after their infection with conditioned medium from RaLV-producing SD-T1 cells. Infectious RaSV was rescued from nonproducer RaSV-transformed Fischer rat embryo cells by pseudotyping the RaSV genome with the Friend murine leukemia virus, and an integrated copy of the RaSV genome was cloned from high-molecular-weight cellular DNA that was obtained from RaSV-transformed NIH 3T3 cells.

Sequencing RaLV and RaSV genomes. The RaLV and RaSV genomes were subcloned prior to sequencing by using standard cloning techniques to generate several smaller DNA fragments that were then used to create sets of unidirectional nested deletions by exonuclease III and S1 nuclease (20). The resultant linear deletion products were recircularized by using T4 DNA ligase and then transformed into an Escherichia coli host in accordance with established procedures (33). Bacterial colonies were randomly picked and propagated, and plasmid DNA was harvested by an alkaline lysis procedure established in our laboratory (28). The sizes of the viral inserts of each of the resultant plasmids were estimated by using restriction enzyme mapping techniques, and DNA fragments containing progressive 170- to 250-bp deletions were denatured at 85°C and subsequently used for DNA sequencing by the dideoxynucleotide chain termination method with a Pharmacia T7 DNA polymerase-based sequencing kit. Reactions that led to ambiguous sequencing results were repeated, with dGTP replaced by d-7-deaza-GTP (d-7dGTP) (36). In these instances, the labeling reactions were performed in an ice bath and the reactions were terminated at 42°C (46). Both DNA strands of each viral insert were sequenced: one was sequenced by using vector-based primers, and its complement was sequenced by utilizing a primer-walking strategy. All oligonucleotide primers were synthesized with a Gene Assembler Plus DNA synthesizer (Pharmacia).

Data entry and analysis. Several programs within the Wisconsin Package (Genetics Computer Group [GCC], Madison, Wis.), including SeqEd, Assembly, Gap, Fetch, LineUp, Names, BestFit, Reverse, Map, PileUp, Distances, GrowTree, and FindPatterns, were used to maintain, manipulate, and analyze the RaLV and RaSV sequence data.

Nucleotide sequence accession numbers. The complete DNA nucleotide sequences of the RaSV and RaLV genomes have been deposited in GenBank under accession nos. M77193 and M77194, respectively.

RESULTS

Genomic structure of RaLV. The complete, primary nucleotide sequence of a single, unintegrated RaLV DNA clone 8.2 kb in length (16) was determined. The genetic structure of RaLV was ascertained by aligning its nucleotides with complete genomic sequences of other mammalian C-type retroviruses, such as MuLV and FeLV, that could be accessed in the GenBank database.

The unintegrated RaLV DNA was composed of 8,107 bp, and the overall structure of RaLV resembled that of other simple, replication-competent retroviruses (Fig. 1). Three long open reading frames (ORF) and one complete long terminal repeat (LTR) could be identified in the RaLV genome. A single SacII restriction endonuclease site, which was used to initially clone this viral DNA, was located between the single LTR and the first long ORF. This evidence supports the notion that this clone of RaLV DNA originated from a one-LTR, circular, viral DNA intermediate, which is commonly produced in acutely infected cells shortly after infection with retroviruses.

The first major ORF of the RaLV genome had the potential to encode a 517-amino-acid-residue precursor Gag polyprotein that was estimated to have a molecular mass of at least 57 kDa (p57\textsuperscript{Gag}). The potential polypeptides that could be encoded by this gene were, in 5’ to 3’ order, a p15 matrix (MA), p12, a p27 capsid (CA), and a p10 nucleocapsid (NC). The predicted numbers of amino acid residues in each of these Gag polypeptides were 128 in the p15 protein, 83 in p12, 248 in p27, and 58 in the p10 protein (Fig. 1).

Two types of signal sequences involved in posttranslational modification of proteins were identified in the MA domain of the RaLV genome. A glycine residue immediately followed the N-terminal methionine of the precursor Gag polyprotein, and this arrangement commonly formed an N-myristoylation site. These data were consistent with the observation that the MA protein of RaLV was myristoylated (4). The second posttranslational signal sequence consisted of two potential N-linked glycosylation sites (N-X-T/S) (2): one was located between amino acid residues 53 and 55, and the other was located between residues 57 and 59. An additional N-linked glycosylation signal sequence was also identified in the CA protein (p27) at amino acid residues 387 to 390; however, the p27 capsid protein isolated from the RaLV particles does not appear to be glycosylated (29).

The deduced NC protein (p10) showed a single zinc-binding domain [C(X)\textsubscript{2}C(X)\textsubscript{4}H(X)\textsubscript{4}C] which is known as a CCHC-type “zinc knuckle,” an element that has been shown to be necessary for RNA-RNA and RNA-protein interactions in other retroviruses (10, 19). A point mutation at nucleotide (nt) 2556 (G to A) was found in the original RaLV clone, and this would be predicted to change the second cysteine residue (TGC codon) of the zinc knuckle to a tyrosine residue (TAC), A
tyrosine kinase phosphorylation site was also found in the NC domain of RaLV at amino acid residues 480 to 488 (KLDKDQCAY) of the Gag polyprotein, which was consistent with the previous observations that this protein was phosphorylated (37).

Identifiable RaLV Pol domains occurred in the 5' to 3' order of protease (PR), reverse transcriptase (RT), RNase, and integrase (IN) (Fig. 1). The numbers of amino acid residues predicted for each of the Pol polypeptides were 187 (PR), 251 (RT), 156 (RNase H), and 362 (IN). The RaLV PR contained

FIG. 1. Nucleotide sequence alignment of RaLV and RaSV. The complete nucleotide sequence of the RaLV genome and the 5' end of the RaSV genome is shown together with the deduced amino acid sequences of the RaLV Gag, Pol, and Env proteins and the RaSV p29 oncoprotein (v-Ra-Ras). The RaSV sequence was aligned to RaLV by the Gap (GCG) program. Dots represent gaps which were inserted into both sequences to maintain the alignment, and vertical lines between the sequences represent agreement between the two genomes. The cloned RaLV genome contained 8,107 bp and one LTR. To properly display the alignment between the RaLV and RaSV genomes, a deduced 8,695-bp clone RaLV provirus with two LTRs is displayed. The boundaries of various regions of viral genomes are indicated by arrows. The nucleotides comprising the 5' and 3' RaLV-ras recombination sites are depicted in boldface, italic letters, and the sequence of the RS-U5 element in the LTR is in boldface type. Other notable LTR elements that are demarcated include positions of the CCAAT and TATAA promoter motifs located within the U3 region, the polyadenylation signal sequence within the R region, direct repeats (DR1 to DR4) and the flanking indirect repeats (IR). The primer binding sites for both positive- and negative-strand DNA synthesis are marked as are putative major 5' and 3' splice sites for env mRNA expression. The SacI and EcoRI restriction endonuclease sites were utilized to initially clone the RaLV and RaSV genomes, respectively. Labeled in boldface letters is the myristoylation signal sequence at the second amino acid residue (Gly) of the p15na gag MA and p29 v-Ra-Ras proteins. The single-amino-acid-residue differences between the amino termini of these two proteins are underscored and in boldface, italic letters.
a 3-amino-acid motif (DTG) encoded by nt 2734 to 2742 that is located in the active site of retroviral proteases. This motif is highly conserved among the PRs of different members of the retrovirus family (52). The putative RT domain of RaLV exhibited a 17-amino-acid motif encoded by nt 3664 to 3714, TLILLQFVDDLLLGATS, containing a pair of Asp-Asp (DD) residues flanked by hydrophobic domains composed of 7 to 8 amino acid residues each. This motif is highly conserved among the retroviral RTs as well as among the RNA-dependent polymerases of many plant, animal, and bacterial viruses, and it is believed to contain the active and/or recognition sites for these polymerases (25). A 194-amino-acid-residue domain, located between the RT and RNase H sequences, was a potential tether that linked these two domains together (23). Located between the RNase H and IN domains of the RaLV genome was a region comprised of 54 amino acid residues (nt 5020 to 5181) that had sequence similarity to pseudoproteases which belong to the dUTPase family and are encoded by poxviruses and some retroviruses (35). A point mutation (G to A) was also found in the pol gene of the RaLV proviral DNA clone at nt 4181. Close to the amino terminus of the RaLV IN domain was a zinc-finger-like HHCC motif composed of a pair of Asp-Asp (DD) residues flanked by hydrophobic domains composed of 7 to 8 amino acid residues each. This motif is highly conserved among the retrovirals as well as among the RNA-dependent polymerases of many plant, animal, and bacterial viruses, and it is believed to contain the active and/or recognition sites for these polymerases (25). A 194-amino-acid-residue domain, located between the RT and RNase H sequences, was a potential tether that linked these two domains together (23). Located between the RNase H and IN domains of the RaLV genome was a region comprised of 54 amino acid residues (nt 5020 to 5181) that had sequence similarity to pseudoproteases which belong to the dUTPase family and are encoded by poxviruses and some retroviruses (35). A point mutation (G to A) was also found in the pol gene of the RaLV proviral DNA clone at nt 4181. Close to the amino terminus of the RaLV IN domain was a zinc-finger-like HHCC motif composed of a pair of Asp-Asp (DD) residues flanked by hydrophobic domains composed of 7 to 8 amino acid residues each. This motif is highly conserved among the retrovirals as well as among the RNA-dependent polymerases of many plant, animal, and bacterial viruses, and it is believed to contain the active and/or recognition sites for these polymerases (25).
of histidine residues (encoded by nt 5203 to 5205) and a pair of cysteine residues (encoded by nt 5314 to 5316) that were separated by 32 amino acid residues, H(X)₃H(X)₃₂C(X)₂C. This motif is strongly conserved in all retroviral and retrotransposon INs in the general form of H(X)₃–7HX₂₃–₃₂CX₂C (52). A second motif defined by a set of 3 amino acid residues with stereotyped spacing that is highly conserved among retroviral INs and transposons is the catalytic core (DX₃₉–₅₈DX₃₅E) of this enzyme (52). A similar motif (DX₃₉DX₃₅E; spanning nt 5464 to 5694) was found in the RaLV IN protein.

The RaLV envelope (env) gene was located at the extreme 3' end of the genome. A major ORF composed of 1,842 bp spanned 6210 to 8051, and its 5' end overlapped the 3' end of the pol gene by 58 nt (Fig. 1). A potential 3' splice site, CATCTTCAG/G (47), was located immediately upstream of this ORF within the pol gene, and this may be utilized in conjunction with a putative major 5' splice site (9) located at nt 651 to 657 in the RaLV genome to generate the subgenomic mRNA(s) that would encode the viral Env polyproteins.

The RaLV envelope (env) gene was located at the extreme 3' end of the genome. A major ORF composed of 1,842 bp spanned 6210 to 8051, and its 3' end overlapped the 3' end of the pol gene by 58 nt (Fig. 1). A potential 3' splice site, CATCTTCAG/G (47), was located immediately upstream of this ORF within the pol gene, and this may be utilized in conjunction with a putative major 5' splice site (9) located at nt 651 to 657 in the RaLV genome to generate the subgenomic mRNA(s) that would encode the viral Env polyproteins.

The Env ORF encoded a potential 614-amino-acid-residue polyprotein.
precursor polyprotein that had a calculated molecular mass of approximately 68 kDa (Pr68env). Located at the amino terminus of the predicted Env protein was a 28-amino-acid-residue leader peptide encoded by nt 6210 to 6293 that is highly conserved among the Env proteins of other mammalian retroviruses. A potential endoprotease cleavage signal site was located between amino acid residues 415 and 418 (RFRR) (encoded by nt 7452 to 7463), and cleavage of the Env precursor protein at this site would be expected to result in a surface protein (SU) composed of 390 amino acid residues (encoded by nt 6294 to 7463) and a transmembrane protein of 196 residues (encoded by nt 7464 to 8051).

The RaLV Env protein displayed 18 cysteine residues that are conserved among the Env proteins of most retroviruses. However, the sixth conserved cysteine residue located in the SU domain was missing and an extra cysteine residue appeared between the conserved first and second cysteine residues. Computer-assisted analysis of the deduced Env polypeptide predicted the presence of two potential membrane-spanning segments within the transmembrane (TM) domain, one spanning amino acid residues 421 to 442 and the other spanning amino acid residues 556 to 581. Within the second potential membrane-spanning domain was a stretch of residues displaying a high degree of hydrophobicity. Additionally, a domain comprised of 45 to 50 amino acids that was predicted to form coiled-coil structures (31) was also located between the two potential membrane-spanning segments.

Five N-linked glycosylation signal sequences (N-X-T/S) (2) were present in the SU region of the RaLV Env polypeptide (Fig. 1), and these were located at amino acid residues 12 to 14, 231 to 233, 275 to 277, 304 to 306, and 307 to 309 of the deduced SU protein. No N-linked glycosylation signal sequences were located within the TM domain.

A purine-rich stretch of nucleotides (5'-AAGAAAATGGGGGGA-3') which spanned the region between nt 8091 and 8103 was located downstream of the presumptive Env stop codon and immediately upstream of the 3'9 LTR. Because of its genomic location and sequence, this element was predicted to act as the primer binding site for positive-strand DNA synthesis (Fig. 1) (13).

Structure of RaLV LTR. The RaLV LTR contained 588 bp, including 438 bp in the U3 region, 72 bp in the R region, and 78 bp in the U5 region (Fig. 1). There were four sets of nearly perfect direct repeat sequences, designated DR1, -2, -3, and -4, in the RaLV LTR (Fig. 1). Each set was composed of two members that differed in sequence from its counterpart within the set by only 1 nt. The DR1 repeats had 10 nt each: one copy (AAAAGACCCA) was found at the 5'9 end of the U3 region, and the other (AAAAGCCCCA) was found at the 3'9 end. The DR2 repeats (CCCAGCCCCC and CCCAGCCCGC) were spaced 14 bp apart from one another within the U3 region, and like the DR1 repeats, they were composed of 10 nt. The DR3 repeats were unique: one was located in the U3 region (GCCCGCACGTAC), while the other was located in the R region (GCCCGCAGGTAC), and each member of this set was composed of 12 nt. The U3 copy of this repeat overlapped one of the DR2 repeats (GCCATCGGTTC and GCTTCTGCTTC) were like the DR2 repeats in that they also were located close to one another (10-bp separation), but they consisted of 11 nt each. The sequences of all four direct repeats of the RaLV LTR were not related to any other known gene or sequences that have been deposited in the DNA sequence databases available to the public. A pair of inverted repeats were located at the extreme 5’ end of the U3 (5'-AA TGGAAAAGACC-3’; underlined sequences lie outside of the
LTR) and the extreme 3' end of the U5 region (5'-GGTCTT TTCATT-3').

The U3 regions of retroviral LTRs include cis-acting control elements that regulate transcriptional initiation of genes by cellular RNA polymerase II. Computer-assisted analysis of this region in RaLV revealed potential binding sites for a number of different eukaryotic transcription factors. Some of these factors are present in many different cell types, such as AP-1 (two potential sites in RaLV U3), AP-2 (three sites), and Sp1 (two sites), whereas others are cell type specific, such as WAP US6 (one site), which has been found in mammary gland cells of lactating mice (30). Three sequence motifs with similarity to the leukemia virus factor a and c (LVa and Lvc) elements present in the Mo-MuLV enhancer were also identified in the RaLV LTR. Factors which bind to these elements in Mo-MuLV have been associated with the pathogenesis of this virus (50). A potential serum response element as well as a binding site for the glucocorticoid receptor could be located at the 5' end of the RaLV U3 sequences. Additionally, three CCAAT sequence motifs and a single TATAA box motif were found in the U3 region (Fig. 1).

Other notable elements found within the RaLV LTR included a polyadenylation signal sequence (AATAAAA) within the R region and a CA dinucleotide which defined the boundary between the R and U5 regions of the LTR.

Four short ORFs that could potentially encode polypeptides 6 to 15 amino acids in length were also located within the U3 region of the RaLV LTR, and a fifth ORF with the potential to encode a 52-amino-acid protein extended from U3 into R (data not shown). A search of the protein databases revealed that these putative proteins do not reveal significant resemblance to any known protein.

The RaLV leader sequences. The 5' untranslated portion of the RaLV genome is the leader sequence which extended 481 nt from the 3' end of the U5 region of the LTR to the initiation codon for the Gag polypeptide (Fig. 1). Located immediately downstream of the 5' LTR was a stretch of 18 nt (5'-TGGG GGCTCGTCCGGGAT-3') that was complementary in sequence to the 3'-most nucleotides of a tRNA\textsuperscript{3'\textprime} molecule, the presumptive primer used by RaLV to initiate reverse transcription (53). Located downstream of this element between nt 651 and 657 was the sequence AGGTAAAG\textsuperscript{C}, which displayed significant sequence identity with that of 5' splice site signal sequences [AGGT(A/G)AGT] (47). Within the same proximity as the putative major 5' splice donor was a purine-rich sequence (5'-GGTTAACCCCGCCCGCCCA-3'), located at nt 652 to 666, that was very similar in sequence (GG CAAAGCCGGCGCGCG) to an element in Ha-MSV that has been shown to be an important component of an RNA dimerization sequence (54). Immediately downstream of this element in both RaLV and RaSV were two copies of a pyrimidine-rich sequence repeat (5'-TGTCTTGTCT-3') at nt 685 to 693 and 695 to 703 that had been shown to also participate in RNA dimerization in Ha-MSV (54). In addition, a third element with a similar motif, which differed by only 1 nt (TGTCT GTGTC; spanning nt 670 to 678), was found between the purine-rich dimerization signal and the tandem pyrimidine-rich elements. Unlike Ha-MSV, these putative dimerization elements did not appear to be related to the rat VL30 (virus-like 30S RNA) elements as determined by BLAST analysis (1). However, a 72-bp stretch of nucleotides (nt 815 to 886) located within the RaLV leader sequence approximately 110 bp downstream of this region and 200 bp upstream of the Gag start codon displayed 95% identity (1% gaps) with the sequence of a rat VL30 element (GenBank accession no. M91235) (data not shown).
mology domains between these two sequences could be identified (data not shown). The first domain, located at the extreme 5′ end of the RaLV U3 region (the first 94 bp), had very little sequence identity with its NICER element counterpart; however, the second domain, composed of 114 bp, displayed 93% identity, with no gaps, with the NICER element. The third domain, of 49 bp, was unique to RaLV, and it appeared as an insert within the NICER element sequence. The fourth domain, of 331 bp, comprising the 3′ end of the U3 region as well as the complete R and U5 regions of the RaLV LTR, displayed 94% sequence identity with 2% gaps between the two sequences.

A search of the nucleic acid databases (Gapped BLASTN, version 2.0.4) also revealed that the 5′ end of the RaLV genome had sequence identity to a rat membrane- and microfilament-associated p58 mRNA (GenBank accession no. S70214 and U15425) (24). This transcript had 97% sequence identity (1% gaps) with the RaLV R, U5, leader, and Gag coding regions (data not shown). An alignment of these two sequences revealed that the p58 mRNA had the potential to encode complete RaLV p15 (MA) and p12 proteins and partial p27 (CA) proteins. Despite the high degree of sequence similarity between the RaLV and p58 sequences, several differences between these two sequences were identified. First, the p58 mRNA contains two gaps of 10 and 7 nt in its leader sequence, in contrast to the RaLV sequence. Second, and most important, the p58 mRNA also contains (in contrast to RaLV) several single-nucleotide deletions and insertions within its gag ORF. Since these apparent deletions and insertions occur within the coding region of the p58 mRNA, the predicted sequence of its protein product is different from that of the RaLV Gag-encoded proteins.

Two additional rat genes which had sequence identities with RaLV were also found by the BLAST analysis, but their sequence similarities were not as extensive as those found for the NICER elements and the p58 mRNA. The rat hepatic steroid hydroxylase IIA2 gene (GenBank accession no. M33313) (34) displayed 95% sequence identity with partial U3 and complete R regions of the RaLV LTR, and the second rat gene was a retrovirus-like, ovary-specific transcript (GenBank accession no. U48828) (15) which had 93% sequence identity to portions of the RaLV leader sequence (5′ end and central region containing homology to the rat VL30 element) and the 3′ end of the LTR U3 region (data not shown).

**Comparison of the RaLV genome with RaSV.** Nucleotide sequence analysis of a 6.8-kb integrated RaSV DNA clone containing the v-Ra-ras oncogene revealed that the RaSV genome comprised 4,479 bp and the remainder of the DNA was derived from the host cellular DNA. The 3′ end of the RaSV sequences ended at an EcoRI site, which was utilized in the original cloning of the viral DNA, close to the end of the RaLV pol gene (Fig. 1).

Heteroduplex mapping studies had shown that (i) a major portion of the v-Ra-ras gene was homologous to and collinear with the p21 ras gene of Ha-MSV and (ii) the c-ras gene had inserted into the RaLV genome near the 5′ end of the gag gene with a concomitant small deletion of RaLV genetic information from this same region of the genome (16). To precisely identify where the v-Ra-ras gene had been inserted into the RaLV genome, the complete nucleotide sequence of RaLV was aligned with the nucleotide sequences of the RaSV genome (Fig. 1). The two genomes were virtually collinear with one another, and the only major difference between the two genomes was the insertion of a 660-bp fragment, the rat c-ras proto-oncogene, into the 5′ end of the MA coding region of the RaLV gag gene. The alignment also revealed that RaSV was missing 88 bp of the MA downstream of the v-Ra-ras oncogene. The sequences flanking the RaSV oncogene (1,167 bp upstream and 2,652 bp downstream) displayed almost complete identity to the RaLV genome, with only sporadic, single-nucleotide changes. Compared to the RaLV genome, 20 bp, including a copy of the inverted repeat element, was missing from the extreme 5′ end of the RaSV LTR.

A single ORF 747 bp in length was located at the 5′ end of the RaSV genome, and this encoded the v-Ra-ras oncogene. This gene was comprised of three distinct domains: two domains were derived from the RaLV genome, and the third was derived from the rat v-Ha-ras-related oncogene. The first 92 bp of the RaLV MA gene comprised the first domain of this ORF. The second domain consisted of a 20-bp element that was derived from the 3′ end of the RaLV LTR (U5), and this element has been designated RS-U5 for RaLV-specific U5 (see below). The third domain was comprised of 635 bp that was virtually identical in sequence to the highly spliced rat c-Ha-ras mRNA expressed in some normal rat cell lines (11).

**Structure of the v-Ra-Ras transforming protein.** The v-Ra-ras oncogene was capable of encoding a protein comprised of 248 amino acid residues with a predicted molecular mass of 27 kDa. This was close in size to a protein with a molecular mass of 29 kDa that has been detected in immunoprecipitates of RaSV-infected cells by using either RaLV Gag-reactive sera or sera reactive to Ha-MuSV p21 (56, 57). These studies also indicated that the RaSV-induced oncogene was a fusion polypeptide consisting of domains encoded by both the RaLV gag gene and the ras cellular proto-oncogene. Alignment of RaSV and RaLVs sequences confirmed these interpretations. The v-Ra-Ras and the RaLV MA proteins had virtually identical N termini (only 1 residue difference among 31 amino acids), and the C terminus of v-Ra-Ras was derived entirely from c-Ha-ras. A tether region linking the N terminus of the RaLV MA protein to the c-ras proto-oncogene in v-Ra-Ras was composed of 7 residues encoded by the RS-U5 element of the RaLV LTR and 21 residues encoded by the first two putative exons (X1 and X2) of the 5′-noncoding region of the c-Ha-ras mRNA (Fig. 3). Although exon 3 of v-Ra-ras encoded the same G12R mutation that was present in v-Ha-ras, the A59T mutation present in v-Ha-ras was not present in the respective region of v-Ra-ras. The fifth c-ras exon incorporated by RaSV and Ha-MSV was similar in both viruses, with the exception of two mutations, R123H and Q150P, in this region of the v-Ra-ras protein. The exon 6 regions of c-ras in both Ha-MSV and RaSV were identical (Fig. 3).

**Identification of the 5′ and 3′ recombination sites between RaLV and the ras proto-oncogene.** The nucleotide sequences surrounding the v-Ra-ras oncogene were analyzed in detail to identify RaLV genomic elements that may have played a role in capturing the c-ras proto-oncogene to form RaSV. Located close to the 5′ end of the v-Ra-ras oncogene within the RaSV genome was the 20 bp RS-U5 element (5′-GAAAGTCTTC TCCGAGGT-3′) which was also located at the 3′ end of the U5 region of the LTR. The first six nucleotides of this element (5′-GAAAGT-3′) were also found at the 5′ end of the gene encoding the RaLV MA protein, while the 3′ end of the RS-U5 element showed sequence identity (10 of 16 nt) with the 5′-noncoding region (X1) of the rat c-ras proto-oncogene (Fig. 4). Downstream of the v-Ra-ras oncogene in the RaSV genome was an 11-bp element (5′-CAGGTAAAGGAGA-3′) that displayed 100% sequence identity with the RaLV MA encoding sequence, at a point located 88 bp downstream of the oncogene insertion site in this gene; however, this element displayed identity with only 7 of 11 nt with a motif located...
downstream of the coding domain at the 3’ end of the c-ras proto-oncogene (Fig. 4).

Sequence similarities between RaLV genes and the c-ras proto-oncogene. Analysis of the RaLV genomic sequences aligned with that of RaSV and the rat c-ras proto-oncogene indicated that small motifs rich in GG, GT, and AG dinucleotides within the 88-bp region of the RaLV-MA gene that had been deleted to accommodate the c-ras gene and generate RaSV exhibited identity with similar motifs within the c-ras proto-oncogene. These motifs may have acted as hot spots for homologous recombination and transduction of c-ras in the form of RaSV.

DISCUSSION

We have identified several novel characteristics of both the RaLV and RaSV genomes that distinguish these viruses from other oncogenic and nononcogenic retroviruses. Whereas RaLV is a replication-competent retrovirus, RaSV is replication defective. This phenomenon is similar to that observed in most acutely transforming viruses which have lost sequences that are essential for their replication while concomitantly incorporating activated proto-oncogenes from tumor cells in which they have replicated. The RaSV-transformed cells express a protein with a molecular mass of 29 kDa that is not present in RaLV-infected cells or any other rat tumor cells used for cocultivation (56, 57). This protein has been shown to be encoded by a chimeric gene consisting of a portion of the RaLV p15<sup>env</sup> gene (MA), a 20-bp element from the RaLV LTR, and the region of c-ras which encodes a 21-kDa protein in Ha-MSV-transformed cells (40, 57).

The significance of RaLV in the transduction of an activated ras proto-oncogene from the rat genome to form RaSV was recognized in our earlier studies, which had indicated that RaLV was the only mammalian retrovirus capable of transducing this gene in vitro when cocultivated with malignant rat tumor cells (44). Analysis of the nucleotide sequences of the RaLV genome revealed that its genomic structure is consistent with that of other simple, replication-competent retroviruses in that it possesses three long ORFs that encode the Gag, Pol, and Env proteins. However, the 8,107-bp RaLV clone sequenced in this work had only one LTR. RaLV is phylogenetically closer to the FeLV group of retroviruses than to various MuLV strains (Fig. 2).

The RaLV LTR is especially noteworthy because it contains a biologically significant motif known as a NICER element (8). The entire R and U5 regions of the RaLV LTR and over half of the U3 region have extensive sequence identity with the NICER elements. The expression of this element in rat adrenal pheochromocytoma cells has been shown to be induced by nerve growth factor and repressed by cAMP. The effects of these agents, if any, upon the expression of RaLV is unknown at this time.

The RaLV mRNA has extensive sequence similarity with a rat membrane- and microfilament-associated p58 mRNA (GenBank accession no. S70214 and U15425). The mRNA...
identity is particularly evident at the 5' end of the RaLV genome, which encompasses the R, U5, leader, and Gag-encoding regions. Two other rat genes which display partial sequence homology with RaLV include the hepatic steroid hydroxylase II-A2 gene (GenBank accession no. M33313) and a retrovirus-like, ovary-specific transcript (GenBank accession no. U48828). These transcripts have partial identity with regions of the RaLV LTR or its leader sequence. Multiple copies of retroviralLTRs and other genes are present in normal cellular DNAs of most vertebrates. Transcription of these elements is affected by numerous environmental stimuli, such as growth factors, hormones, chemical carcinogens, and radiation. It is therefore possible that RaLV-related retrotransposons or other RaLV-related genetic sequences may be transcribed individually to form defective particles, express partial transcripts, or produce complete replication-competent viruses at different stages of normal or abnormal growth of the rat cells.

Nucleotide sequence analysis of the RaSV genome revealed that (i) it was related to RaLV since the two genomes were homologous and largely collinear with one another and (ii) the v-Ra-ras oncogene was inserted into the RaLV genome near the 5' end of the gag gene with a concomitant loss of 88 bp from gag. The only major ORF of the RaSV genome was that for the 747-bp v-Ra-ras oncogene. The sequencing data confirmed our earlier observations that the 5' end of v-Ra-ras was derived from a portion of the RaLV gag gene (MA), and the 3' end was derived from a c-ras proto-oncogene. However, our new data demonstrate that the v-Ra-Ras p29 oncoprotein is encoded by a complex fusion gene in which different regions of the endogenous RaLV genome have been joined to all six putative exons of an activated c-ras gene to form the 747-bp ORF. This recombinational event is novel to the RaSV genome. Patterned boxes represent specific nucleic acid sequence domains from different regions of the RaLV genome or the c-ras proto-oncogene; as in Fig. 3, boxes with the same design or shading represent similar or identical nucleic acid sequences. RaSV-63SP is a second RaSV isolate recovered independently in Japan by in vitro cocultivation of RaLV-productive SD1-T cells with a cell line established from a rat mammary tumor induced by dimethylbenzanthracene (27).

The N-terminal region of v-Ra-Ras is unique among the Ras proteins in that it contains 59 additional amino acid residues (39). Thirty-one of these residues are derived from RaLV MA, 7 are derived from the RaLV RS-U5 element, 7 are derived from exon 2 of c-ras, and 14 are derived from the 5' untranslated end of the third exon. The seven-amino-acid peptide encoded by the RS-U5 element is not found in any other protein whose sequence has been deposited in the public databases.

Two amino acid residue changes in the v-Ha-Ras protein sequence, in contrast to the c-Ras (G12R and A59T), have been shown to be important for its transforming activity (5, 6, 12). In RaSV, the deduced v-Ra-Ras protein has the G12R mutation but it has retained the alanine residue at position 59 (i.e., there is no change from the wild-type c-Ras sequence). How only the G12R and not the A59T mutation in v-Ra-Ras plays a role in cellular transformation is not clear at present. However, it can be speculated that, individually or collectively, (i) the myristoylation and nuclear localization signals present at the amino terminus of the v-Ra-Ras protein, (ii) the amino acid residues encoded by the RS-U5 element of the RaLV LTR, or (iii) the c-ras noncoding exons may cooperate in conferring transformation properties to this protein. It should be noted that transforming properties can be conferred to normal cellular p21 Ras proteins by adding a myristoylation signal sequence to their amino termini (4).

The molecular events that are involved in the biogenesis of
highly oncogenic viruses either in vivo or in vitro are not well understood. A role for VL30 elements in the formation of both the Ha-MSV and Ki-MSV genomes has been suggested since their v-ras genes are bracketed by VL30 sequences (7, 14, 32). Prior to the availability of DNA sequence data for both the RaLV and RaSV genomes, it had been speculated that the VL30 elements play a role in the transduction of c-ras by RaLV (14, 32). Our data clearly demonstrate that the putative VL30-like elements previously noted in the RaSV genome (EMBL accession no. X00840 and GenBank accession no. M91235) are an integral part of the 481-bp leader sequences present in both the RaLV and RaSV genomes. These sequences are identical in replication-competent RaLV as well as in two independent isolates of RaSV, one isolated in the United States (RaSV) (44) and the other isolated in Japan (RaSV-63SP) (27). Moreover, the similarity between the VL30-like elements and the RaLV or RaSV leader sequences appear to end 196 bp upstream of the 5’ RaLV-ras junction in RaSV in a region which contains the putative major 5’ splice site as well as the RNA packaging and dimerization signal sequences of these viruses. Since both the 5’ and 3’ RaLV-ras junction sites in RaSV are derived exclusively from the RaLV MA gene, no VL30-like elements are involved in the recombination between the RaLV and c-ras genes.

Duplication of the 20-bp RS-U5 element from the U5 region of the RaLV LTR is unique to RaSV. Not only does this sequence encode a portion of the transforming v-Ra-Ras p29 protein, but it also appears to be involved in the recombination between the RaLV MA gene and the 5’ end of c-ras (i.e., it represents the 5’ recombination site).

Several mechanisms of both homologous and nonhomologous recombination of retroviral and cellular elements have been proposed (17, 21, 22, 38, 51). Although recombination during reverse transcription and template switching in the generation of RaSV cannot be completely ruled out, it involves formation of a 3’ junction of a cellular gene with a viral gene(s); i.e., it first forms a “hybrid” gene. This hybrid gene (i.e., RaLV-ras) could not be detected in Southern blots of genomic, high-molecular-weight DNA extracted from SD-1T cells and the 4NQO cell line established from rat tumors induced by the chemical carcinogen 4-nitroquinoline N-oxide; i.e., no hybridization of a DNA fragment with both the RaLV and ras probes was detected (unpublished data).

An abundance of GG dinucleotide repeats or GT-rich trinucleotides in the genomic DNA has been shown to increase homologous recombination in both the prokaryotic and eukaryotic genomes (3, 55). In RaSV, we have identified sites containing GG dinucleotide repeats and GGA, GGT, and/or AGG trinucleotides in each of the four or five different viral and cellular sequences that have recombined to generate the v-Ra-ras oncogene. These motifs were found in the coding as well as noncoding exons (X2 and part of X3) of c-ras as well as the RaLV MA and RS-U5 element from the RaLV LTR (Fig. 4).

The RS-U5 element is particularly noteworthy since it contains two distinct hot spots; its 5’ end contains an AGG trinucleotide and a GG dinucleotide, which overlap and recombine with an identical sequence in the RaLV genome exactly where MA breaks to allow insertion of the RS-U5 sequence. The 3’ end of RS-U5 is also rich in GG, GGT, and GGA motifs, which may have facilitated homologous recombination of the RS-U5 element to the c-ras noncoding exons via a GGT-containing overlapping sequence from the first putative exon (X1) of the c-ras proto-oncogene. Exon 1 does not encode a protein but it appears to participate in the recombination of the viral RS-U5 element to c-ras (Fig. 3 and 4). An interesting observation was that 88 bp of the p150N MA gene which were lost during recombination with the c-ras gene also exhibited stretches of hot spots 5 to 10 nt long that are scattered throughout the RaLV MA and c-ras noncoding exons. These data provide an explanation as to why RaLV has recombined consistently with the same c-ras sequences during cocultivation of SD-1T cells with rat tumor cells while cocultivation of the same rat tumor cells separately with several MuLV and FeLV strains does not yield acutely transforming viruses (29, 44). Whatever mechanism underlies RaLV-ras recombination to form the RaSV genome, the pathway involved is likely to be distinct from those hypothesized for the generation of Ha-MSV, Ki-MSV, or other mouse sarcoma viruses.

The rat cell and rat endogenous retrovirus system that we have established offers a unique opportunity to study how endogenous, replication-competent retroviruses or retrotransposon-like elements interact with cells of their hosts. Studies utilizing this homologous system may provide unique insights into how environmental factors affect the expression of endogenous retroelements, how these elements interact with specific cellular genes, and how transcriptional or translational regulation of these genes may influence the course of disease pathogenesis within the host. Recently, we have cloned full-length RaLV proviruses and LTRs from SD-1 and SD-1T cells (20a). These retroelements can now be manipulated to sequentially delineate the molecular steps involved in the expression or suppression of specific genetic sequences that may contribute to the development of cancer and/or other diseases in rats.

ACKNOWLEDGMENTS

We are grateful to Howard Young and Matthew Gonda for their collaborations in earlier studies and for providing the RaLV and RaSV plasmid clones. We also thank Norman Frolich for his assistance with the GCG software package. Grant support to Sauriya Rasheed (CA 27246) from the National Cancer Institute for the earlier studies and support from the Rasheed Research Fund for the more recent work are acknowledged.

REFERENCES


