

Alignment of U3 Region Sequences of Mammalian Type C Viruses: Identification of Highly Conserved Motifs and Implications for Enhancer Design

ERICA A. GOLEMIS,[†] NANCY A. SPECK,[‡] AND NANCY HOPKINS*

Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Received 4 August 1989/Accepted 12 October 1989

We aligned published sequences for the U3 region of 35 type C mammalian retroviruses. The alignment reveals that certain sequence motifs within the U3 region are strikingly conserved. A number of these motifs correspond to previously identified sites. In particular, we found that the enhancer region of most of the viruses examined contains a binding site for leukemia virus factor b, a viral corelike element, the consensus motif for nuclear factor 1, and the glucocorticoid response element. Most viruses containing more than one copy of enhancer sequences include these binding sites in both copies of the repeat. We consider this set of binding sites to constitute a framework for the enhancers of this set of viruses. Other highly conserved motifs in the U3 region include the retrovirus inverted repeat sequence, a negative regulatory element, and the CCAAT and TATA boxes. In addition, we identified two novel motifs in the promoter region that were exceptionally highly conserved but have not been previously described.

Many genetic studies have shown that transcriptional signals in the U3 regions of mouse type C retroviruses, particularly enhancer elements, are potent determinants of viral pathogenicity (9, 10, 18-20, 30-32, 45, 47, 49, 68). In nondefective viruses that induce hematopoietic tumors, transcriptional sequences can influence leukemogenicity, the latency period of disease induction, and disease specificity (9, 10, 18-20, 30-32, 45, 47, 49). The type C mammalian retroviruses constitute one of the best model systems for studying how transcriptional regulation of viral gene expression influences viral pathogenesis. There are more than 100 independent isolates of type C viruses. A significant percentage of these have been molecularly cloned and sequenced, and in a number of cases pathogenic phenotypes have been mapped to the viral enhancer-promoter region (for reviews, see references 82 and 83).

The mouse type C virus enhancer is present as a direct repeat of about 50 to 150 nucleotides; it is located approximately 200 base pairs (bp) upstream of the cap site (28, 42, 43, 46). Workers in our laboratory and others are engaged in detailed genetic and biochemical analyses of murine type C virus enhancers to understand how nuclear factors interact with these elements to shape the disease-inducing phenotypes of the viruses (5, 7, 27, 37, 50, 64, 65, 68, 71, 72; N. A. Speck, B. Renjifo, E. Golemis, T. N. Fredrickson, J. W. Hartley, and N. Hopkins, *Genes Dev.*, in press). For this purpose we found it useful to prepare an alignment of published type C virus enhancer sequences, particularly those of mouse type C viruses whose pathogenicity has been studied. Besides helping us to begin correlating specific enhancer factor-binding sites with pathogenic phenotypes, particularly disease specificity, this alignment led to the observation of a highly conserved structure to the enhancer elements of this particular group of mammalian type C

viruses. Encouraged by these observations, we proceeded to align the rest of the U3 region sequences to search for additional motifs whose unusually high conservation might implicate them as important *cis*-acting elements in viral replication. 5' of the enhancer region, two previously defined motifs, one involved in integration and the other an apparent negative regulatory element, stood out as being particularly highly conserved. In the promoter region, in addition to the CCAAT and TATA sequences, two blocks of extremely highly conserved sequence of unknown function were noted between the CCAAT and TATA motifs. In this paper, we provide the sequence alignments that have proved useful in our own research and describe highly conserved aspects of the enhancers of these particular mammalian type C retroviruses.

METHODOLOGY AND RESULTS

Rationale of the data base. The U3 regions in type C murine retroviruses are between 350 and 550 bp in length. We compiled a data base that contained complete U3 region sequences of 32 different type C mammalian retroviruses and partial U3 sequences spanning the enhancer region and its flanking sequences for three additional retroviruses (the complete alignment is available upon request). All sequences used in this study were either directly transferred from the GenBank database or obtained from the references indicated in the legend to Fig. 2. Alignment was assisted by using the Multiple Sequence Editor program of the University of Wisconsin Genetics Computer Group (22).

All of the viruses included in this comparison were derived either from cloned replication intermediates or from full-length integrated proviruses of ecotropic, xenotropic, amphotropic, and polytropic host range classes. We excluded sequences for truncated endogenous proviral fragments, since sequence variation in the U3 region of these viruses may have occurred in the absence of selective pressure for viral viability. Although most of the viruses in the alignment are murine leukemia viruses (MLVs), several feline and simian type C viruses were also included.

Some of the viruses included in the alignment are more

* Corresponding author.

[†] Present address: Department of Molecular Biology, Massachusetts General Hospital, 50 Blossom St., Boston, MA 02114.

[‡] Present address: Department of Biochemistry, Dartmouth Medical School, Hanover, NH 03756.

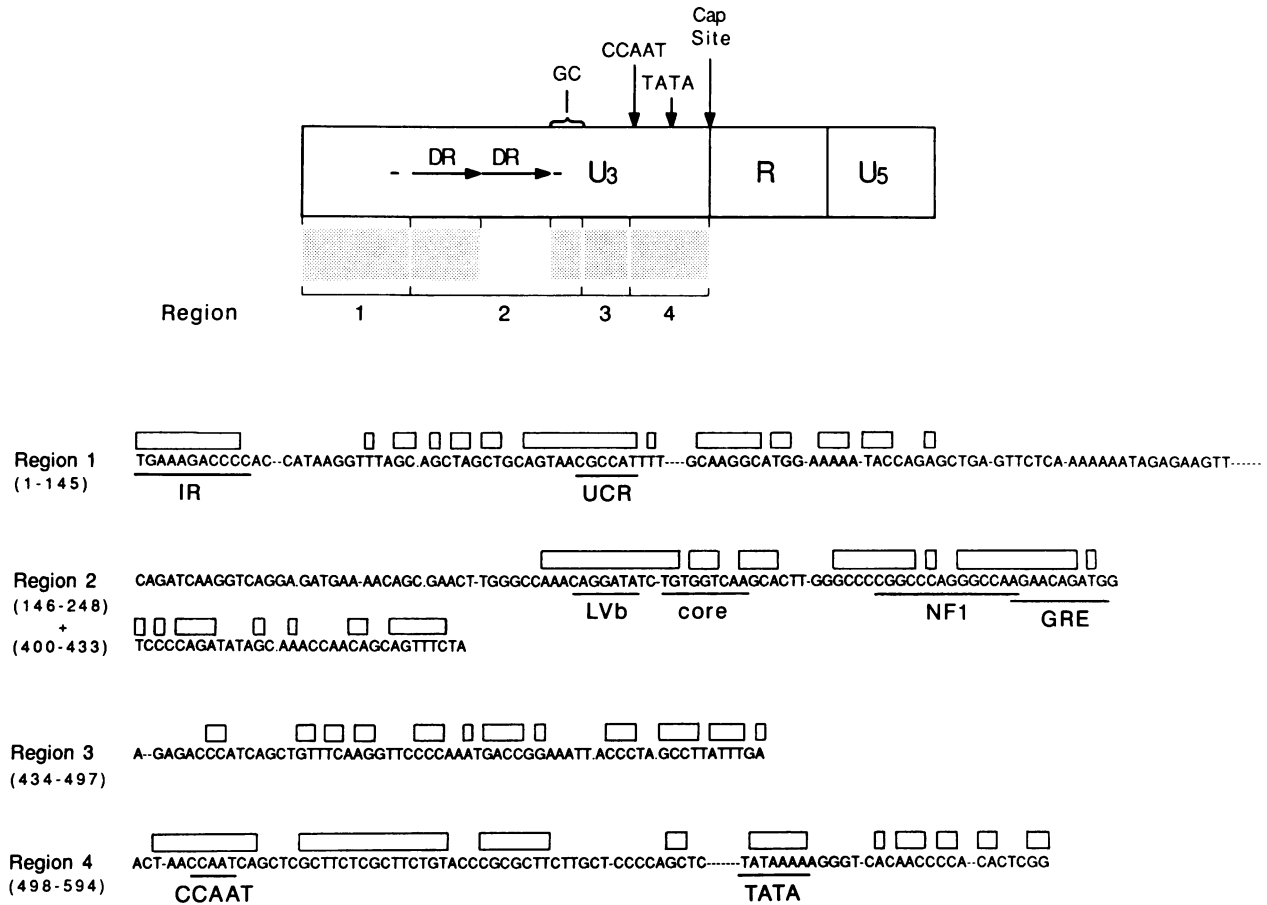


FIG. 1. Plurality sequence for the U3 region of 35 mammalian type C viruses indicating sequences shared by >90% of the isolates. The top part of the figure is a schematic representation of the LTR, showing the location within U3 of the four regions described in the text. Viruses whose sequences were aligned to generate the plurality sequence are described in the legend to Fig. 2. The plurality for the aligned sequences was created by using the Multiple Sequence Editor program and is represented for each of regions 1, 2, 3, and 4 on the bottom part of the figure. Symbols: . , two different bases are equally prevalent at that position; -, most viruses do not contain a nucleotide at that position, but that it was necessary to introduce a gap to maximally align some of the viruses. For each base position, the number of viruses matching the plurality was calculated and divided by the total number of viruses compared for that position. (In practice, to achieve a conservation score of 90% or higher, no more than three viruses diverged at any given position.) If two different bases were equally prevalent at a position, one was arbitrarily defined as divergent and viruses with that base were scored as varying from the plurality. In cases in which individual viruses had point insertions, the virus with the insertion was defined as varying from the plurality. Any base in the plurality that is shared by more than 90% of the viruses is indicated by an open box drawn above it. The IR sequence required for provirus integration, and highly conserved binding sites for several cellular factors are underlined.

closely related than others (see the legend to Fig. 2). In some cases, one virus gave rise to others through laboratory manipulations. For instance, Moloney murine sarcoma virus (MoMSV), Abelson leukemia virus, and myeloproliferative sarcoma virus are all variants of Moloney MLV (MoMLV), although they were not recovered from a common biologically or molecularly cloned isolate of MoMLV (1, 2, 12, 52). Although there will be less sequence variation between the more closely related isolates, the variability is particularly useful because it should be restricted to nonessential sequences, or, alternatively, may reside in motifs that specify distinct biological properties of the different isolates.

For the purposes of comparison, we divided the U3 region into four segments (Fig. 1). Region 1 consists of about 145 bp at the 5' end of U3. Region 2 consists of approximately 100 bp immediately 3' to region 1 and includes enhancer sequences corresponding to the first copy of the directly repeated sequence for viruses that have such a repeat. An additional 35 nucleotides immediately 3' of the direct repeat

in most viruses but included in the direct repeat of others was also assigned to region 2 (Fig. 2A). Region 3 consists of about 65 bp between region 2 and the CCAAT box, and region 4 consists of 100 bp at the extreme 3' end of U3 that contains the viral promoter (Fig. 1) (28, 56). We determined the base that appears most frequently at each position in the alignment (Fig. 1; Fig. 2A, top line; see also Fig. 4). If that base is shared by >90% of the isolates in the study, we have placed an open box over it in Fig. 1 and drawn a box around it in Fig. 2A and 4.

The percentage of bases matching the plurality sequence is 75% or greater for all four regions; the promoter-proximal sequences (region 4) are the most highly conserved. This degree of conservation clearly reflects the close evolutionary relationship of the type C viruses. Variation is not uniform. In each region, short stretches of very highly conserved sequence are interspersed with regions that are much more variable; this is especially true for the enhancer region, as discussed below. Highly conserved motifs, particularly con-

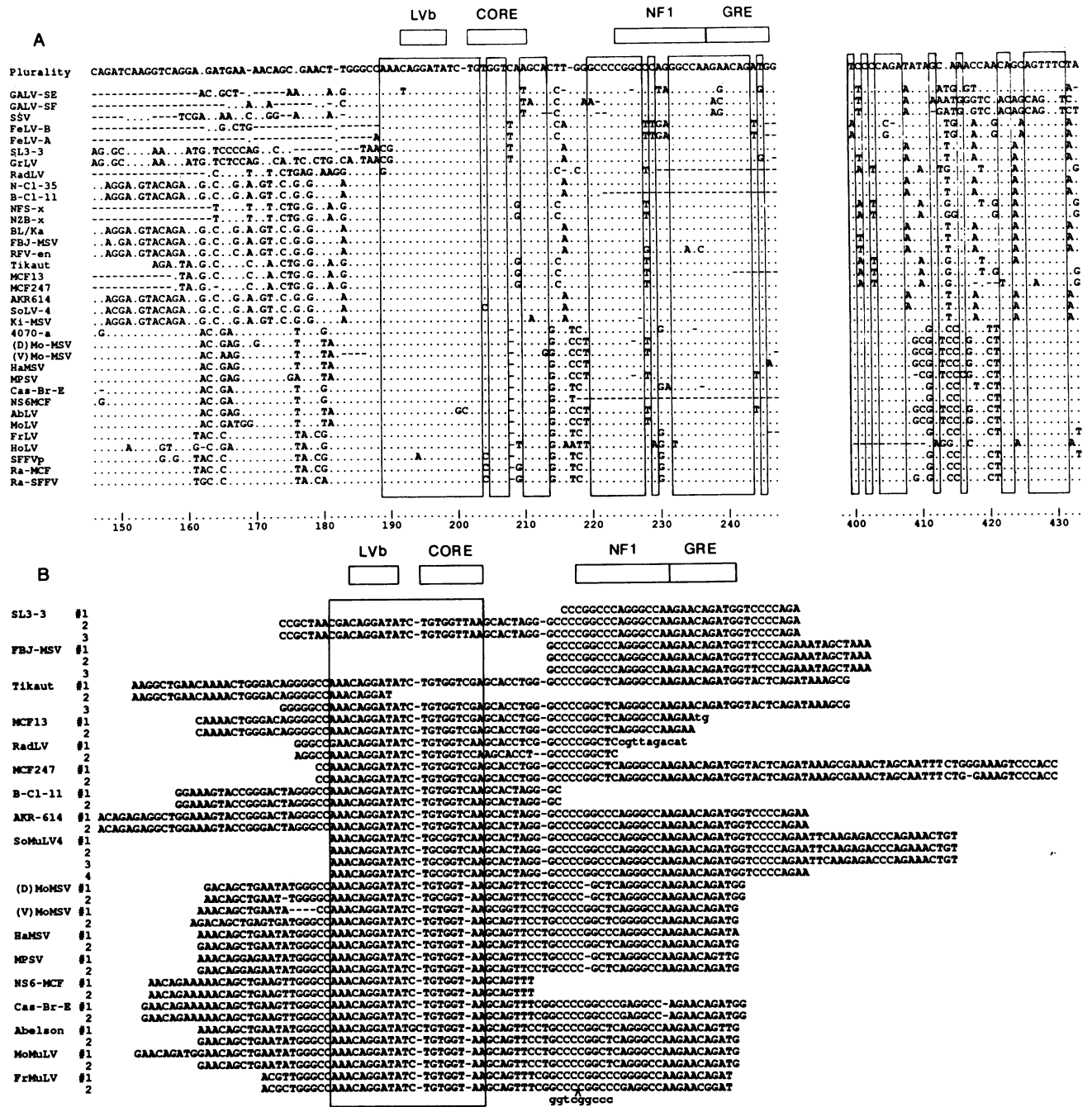


FIG. 2. Alignment of enhancer sequences (region 2 from Fig. 1) of murine, simian, and feline type C retroviruses; delineation of conserved motifs and indication of repeat structure. Viruses included in this alignment are abbreviated as follows: GALV-SE and GALV-SF, gibbon ape leukemia virus, SEATO and San Francisco isolates (73); SSV, simian sarcoma virus (21); FeLV-B and FeLV-A, feline leukemia virus, B subgroup/Gardner-Arnstein, and A subgroup/Glasgow-1 (67); SL3-3, SL3 lymphomagenic virus (45); GrLV, Gross passage A leukemia virus (79); RadLV, RadLV/VL3 (T⁺L⁺) radiation-induced leukemia virus (33); N-CI-35, nonthymotropic murine retrovirus, and B-CI-11, thymotropic murine leukemia virus (19); NFS-x, NFS-Th-1 xenotropic leukemia virus (38); NZB-x, NZB xenotropic murine leukemia virus (55); BL/Ka, nonleukemogenic cloned C57BL/Ka provirus (39); FBJ-MSV, FBJ murine sarcoma virus (77); RFV-en, chemically induced endogenous retrovirus, from RFM-Un mice (48); Tikaut, Tikaut leukemia virus (16); MCF13, mink cell focus-forming virus (MCF) isolate 13 (70); MCF247, MCF isolate 247 (36); AKR614, AKR clone 614 nonleukemogenic virus (76); SoLV-4, Soule leukemia virus, isolate 4 (16); Ki-MSV, Kirsten murine sarcoma virus (54); 4070-a, amphotropic murine retrovirus, clone 4070 (63); (D) Mo-MSV, Moloney murine sarcoma virus (24); (V) Mo-MSV, Moloney murine sarcoma virus (78); HaMSV, Harvey murine sarcoma virus (83); MPSV, myeloproliferative sarcoma virus (66); Cas-Br-E, Lake Casitas Brain E neurotropic retrovirus (35); NS6MCF, MCF derived from Cas-Br-M in NFS mice (11); ABLV, Abelson murine leukemia virus (61); MoLV, Moloney murine leukemia virus (62); FrLV, Friend murine leukemia virus, clone 57 (40); HoLV, Ho wild mouse leukemia virus (80a); Fr-SFFVp, Friend spleen focus-forming virus, polycythemia inducing (13); Ra-MCF, Rauscher MCF (80); Ra-SFFV, Rauscher spleen focus-forming virus (4). (A) Sequences shown comprise the first copy of the direct repeat in most viruses

served blocks at least 5 to 6 bp in length, are candidates for *cis*-acting regulatory sequences for viral gene expression. We suspect that in regions where colinearity among the virus sequences is maintained, the spacing of neighboring motifs may also be functionally important.

Enhancer region. Region 2 contains the retroviral enhancer, which has been considered to be a hypervariable region in U3 (Fig. 2) (14). A significant source of this variability (as well as of variability at the 3' end of region 1) comes from differences in the boundaries of the directly repeated sequence. Naturally occurring retroviruses often have two or more tandem copies of enhancer sequences, but the sequence that is repeated varies between viruses; the location of the 5' boundary of the direct repeat can vary over 51 bases (positions 138 to 189 in our alignment), and the 3' repeat boundary can vary over 75 bases (positions 217 to 292) (Fig. 2B). To identify maximum conservation of the enhancer sequences, it was therefore necessary to introduce gaps in the sequences to obtain the optimal alignment.

Figure 2A shows the alignment of enhancer region sequences. The left-hand side of the figure contains sequences present in the first copy of the direct repeat or, in some cases, lying just 5' of the sequence that is repeated (Fig. 2B). The right-hand side of Fig. 2A shows sequences that are occasionally included within the repeated sequence, but are more often found immediately 3' to the direct repeat and have been shown in some cases to contribute to enhancer-promoter activity (43).

There is striking conservation of the motif AACAGG ATATCTG(T/C)GGT in the viral enhancer, and this sequence is usually flanked on its 5' side by GGGCC (Fig. 2A). The 18-bp stretch is 98% conserved overall, with sequences from 26 of 35 viral isolates matching the plurality exactly. A number of different lines of evidence argue that this motif is a key component of the mammalian type C retroviral enhancer. Of the 18 viral isolates we compared that have enhancer duplications, all but one of the isolates include this motif in the duplicated region (Fig. 2B [some data not included]). A central portion of this motif, the CAGGATA motif, was originally defined by mobility shift assays as the binding site for a ubiquitous factor, leukemia virus factor b (LVb), and subsequently defined as the binding site for a factor designated leukemia virus factor t (LVt) (64; N. R. Manley, M. A. O'Connell, and N. Hopkins, unpublished results). Mutation of two G nucleotides in the LVb/LVt-binding site (to CATTATA) significantly attenuates transcription from the MoMLV enhancer (65). The LVb-binding site is also part of an approximately 30-bp sequence element in the MoMLV enhancer that mediates transcriptional activation by 1,3-phorbol myristate acetate (TPA) in Jurkat T cells (65). Especially in light of this TPA inducibility, it is intriguing that in many viruses the 5' end of the 18-bp conserved motif, along with two C's from the GGGCC motif, include the sequence CCAACAGG, which is similar to the consensus for the CARG box. This CARG motif is found in

the central portion of the serum response element, which is bound by the p67 serum response factor (26, 29, 60, 72, 74, 75). It has not yet been determined whether this homologous CARG motif contributes to TPA inducibility of the retroviral enhancer.

The 3' portion of the highly conserved motif in the direct repeat corresponds to part of the consensus for the conserved viral "core" motif, found also in the simian virus 40 and polyomavirus enhancers [TGTGG(T/A)(T/A)(T/A)] (81). This region is also included in the TPA-responsive element of MoMLV (65). Inspection of the type C retroviral enhancer sequences in Fig. 2A reveals a number of variants of the core sequence, for example, TGTGGTCAA, TGTGGTAA, TGTGGTTA, TGTGGTCGA, TGCGGTGAG, and TGCGGTAA. Proteins have been shown to bind to some of these variants, and studies in several laboratories have implicated certain versions of this motif to be involved in the T-cell- or lymphoid-specific transcriptional preference of the MoMLV and SL3-3 enhancers (5, 34, 64, 65, 72).

Yet another line of evidence that suggests an important role for the LVb/core conserved region comes from the study of the pathogenicity of viruses containing point mutations in either the LVb- or the core-binding sites of the MoMLV enhancer. Although these viruses are viable, they exhibit an increase in the latency period of disease induction relative to the wild-type MoMLV and alterations in the types of leukemia that they induce (Speck et al., submitted).

Sequence conservation throughout the remainder of the enhancer region is somewhat less dramatic. Furthermore, because the repeat structure of different viral isolates causes the 3' boundary of the repeat to vary over 75 bp, it is more difficult to compare sequences quantitatively in this region. However, with the sequences aligned as shown in Fig. 2A, it is possible to identify a second set of motifs that are quite highly conserved, although they are less frequently duplicated than the LVb/core motif. These motifs include the bases GCCCCGGC (96% overall), followed by four variable bases, and then GGCCAAGAACAG (96% overall). These bases include the consensus binding motif for nuclear factor 1 (NF1) (T/CGGN₅₋₆GCCAA) (23, 44, 53). Of 35 viruses, 29 contain the exact binding site for NF1, and binding of purified NF1 to this site has been demonstrated for MoMSV (B. Graves, personal communication). Mutation of the NF1 site results in a decrease in transcriptional activity conferred by the MoMLV enhancer, with the decrease being most pronounced in fibroblast cell lines (65). The significance of the conservation of the GCCC motif flanking the NF1 consensus on its 5' side is not known.

Many viruses also possess the consensus for the glucocorticoid response element (GRE) (AGAACAGATG), a binding site for steroid hormone receptors (7, 17, 51, 84). The GRE mediates the dexamethasone induction of several viruses, including SL3-3, AKV, and MoMSV (7, 17, 51). Twenty-seven viruses maintain the GRE precisely in at least one copy of their enhancer sequence, and in several enhancers

(left side of the figure) and a short sequence that lies 3' of the direct repeat in most viruses (right side of the figure). Any base in the alignment for which more than 90% of the sequences match the plurality sequence is boxed. The repeat structure of a number of the included viruses is such that the 3' end of the enhancer region repeat is "truncated" (panel B), although certain motifs are present in the second copy. Boxes reflecting conservation of sequence are thus drawn to indicate the presence of the conserved sequence in at least one copy of the enhancer sequence, even though the second-copy sequence is not shown. Binding sites for the cellular proteins LVb, core, NF1, and GRE, shown to bind to at least some isolates, are indicated above the alignment. (B) The viruses shown here are a subset of those included in panel A. The entire repeated sequence for each isolate is shown. Where tandem copies of repeat sequence are interrupted by short nonrepeated sequences, the interruptions are indicated in lowercase letters. For Friend virus, a single 9-bp insertion in the middle of the second copy of the enhancer sequences is also indicated in lowercase letters. The region corresponding to the LVb/core conserved domain is boxed in a heavy line; the region corresponding to the NF-1 and GRE consensus motifs is indicated by lines drawn above the aligned sequences.

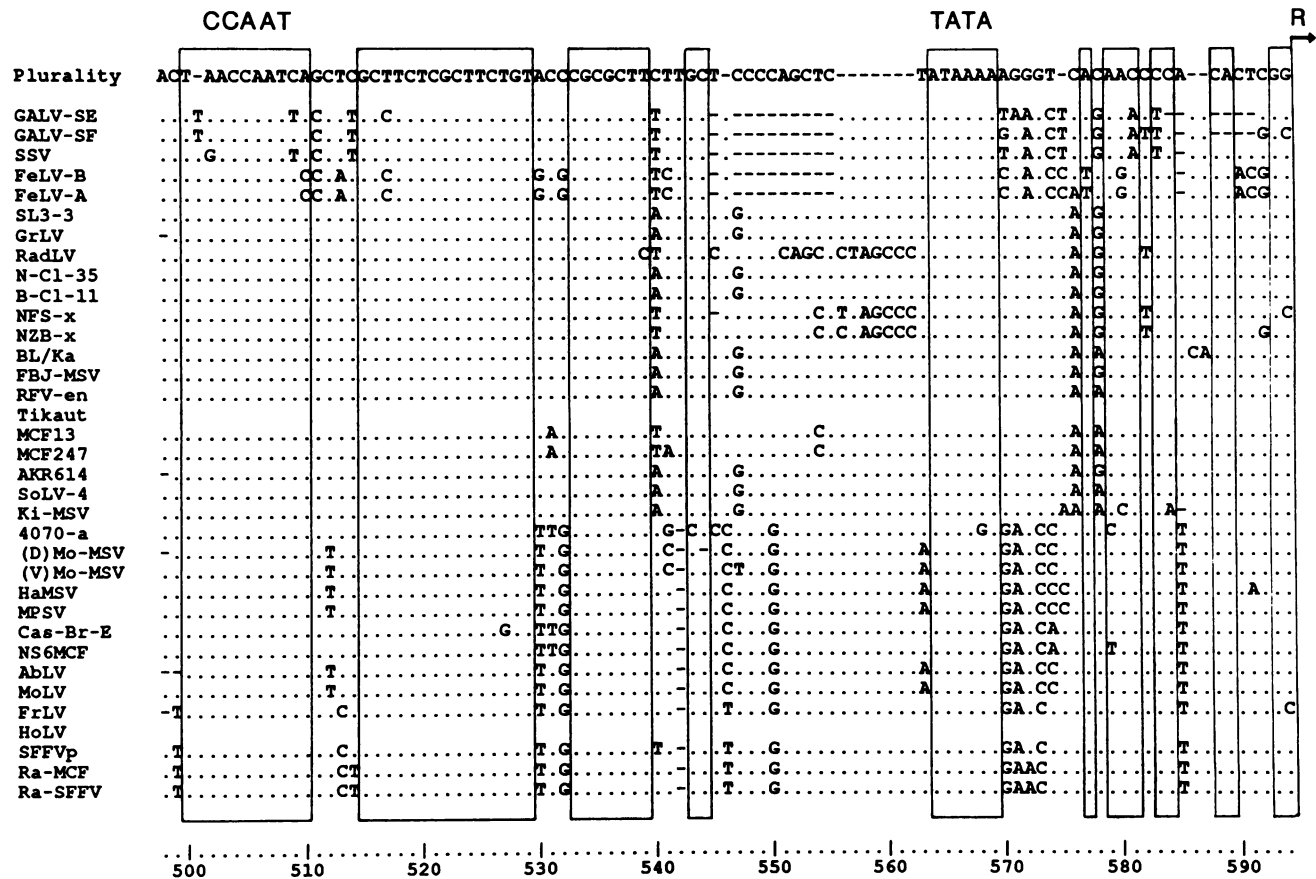


FIG. 3. Alignment of promoter region sequences (Fig. 1, region 4). See the legend to Fig. 2 for a description of viruses whose sequences were aligned and for criteria for drawing boxes around highly conserved bases.

(e.g., MoMLV and SL3-3) the GRE site is repeated three times. Not all of the enhancers that have GRE sites are transcriptionally activated by dexamethasone; this might be due, in part, to flanking sequences that affect productive utilization of the GRE (17, 57).

In the region from the LVb/core motif through the GRE (Fig. 2A, positions 188 to 248), no large gaps are introduced in the sequences to make different viruses fit the alignment. This suggests that not only maintenance of specific sequences, but also their spacing relative to each other, may be important for optimal enhancer function. Most of the viruses compared have all four of these motifs (LVb-core-NF1-GRE), and many include them in their repeat structure.

The right-hand segment of Fig. 2A shows approximately 35 bp of sequences that are within the direct repeat of a few viruses but are most often found flanking the repeat on the 3' side of the promoter-proximal GRE site. Laimins et al. called these sequences in MoMSV the "GC rich" element and determined that they contribute to transcriptional activity (43). There is one short, highly conserved motif (AGT TTC) in the 3' half of the 35-bp segment (Fig. 2A), but its function, if any, has not been determined. The 5' half of this segment differs considerably among different isolates. Consonant with this diversity among isolates, studies by Golemis et al. show that the 5' half of this 35-bp segment contains a determinant of erythroleukemogenicity in the clone 57 isolate of nondefective Friend virus and imply that the corresponding region of Moloney virus may encode a determinant of T cell lymphomagenesis (27). The Moloney and Friend

virus sequences from this region bind distinct nuclear factors, which may contribute to the disease specificity conferred by this region (50).

Conserved sequences at the 5' end of U3. In the 5' region of U3 (region 1) there are two long blocks of highly conserved sequence (Fig. 1). The sequence TGAAAGACCCC (AATG AAAGACCCC before integration) corresponds to the inverted repeat (IR) sequence essential for retrovirus replication (58). Conservation of this consensus is 97%; 31 of 35 viruses maintain the exact sequence. A second long block of conserved sequence includes the motif CGCCAAT, included in the upstream conserved region (UCR) site. Flanagan et al., by a combination of *in vitro* protein binding and *in vivo* functional analyses, recently identified this motif as a negative regulatory element for several murine viruses and noted the striking conservation of this motif among a large number of type C retroviruses (25). Spacing between the IR and UCR sites is somewhat flexible; although their borders are about 27 bp apart in most isolates, in some viruses the distance between them is as short as 18 bp. Besides the IR and the UCR motifs, no other long conserved motifs were found in region 1, although a 6-bp block of highly conserved sequence flanks the 3' side of the UCR motif. As noted above, it is necessary to introduce gaps in the alignment toward the 3' end of region 1 to "fit" different viruses.

Promoter-associated motifs. Region 4, which includes the transcriptional promoter, shows the highest level of conservation in the U3 sequences (Fig. 1 and 3). In particular, the CCAAT and TATA motifs are almost completely conserved

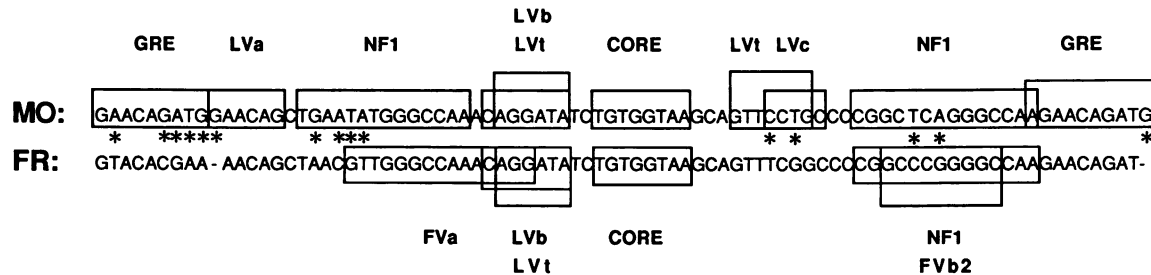


FIG. 4. Binding sites for cellular proteins on the MoMLV and Friend MLV enhancers. Binding sites for nuclear factors on the first copy of the direct repeat (enhancer) sequences of Friend MLV (50; Manley et al., unpublished) and MoMLV (64). Boxes indicate the boundaries of binding sites, usually as defined by methylation interference (50, 64). Binding of glucocorticoid receptor to the Friend MLV enhancer has not yet been studied. It has not been determined which factors can bind simultaneously to these elements.

(6). There is also striking conservation of sequences between the CCAAT and TATAA motifs. In particular, the 15-bp motif GCTTCTCGCTTCTGT is 99.2% conserved and is exactly maintained in 29 of 33 viruses, whereas the motif CGCGCTT is 99.6% conserved and is identical in 32 of 33 viruses. These highly conserved motifs coincide with a region suggested by Graves et al. to be part of a "distal signal" for MSV long terminal repeat (LTR)-directed transcription (28). Although the function of these conserved sequences between CCAAT and TATA has not yet been investigated in detail, they may be important promoter-proximal elements. The spacing of most of these large blocks of conserved sequence is quite rigorously maintained; no gaps at all intervene between CCAAT and either of the two novel downstream conserved motifs. More variable spacing occurs between the CGCGCTT motif and TATA; the distance between these motifs ranges from 5 to 20 bp.

The approximately 65-bp segment in region 3 between the enhancer region (region 2) and the CCAAT box is less highly conserved than the promoter region (Fig. 1). There are no highly conserved motifs longer than 4 bp, and to our knowledge, no function has yet been identified for this region of U3.

DISCUSSION

We have generated a computer alignment of published enhancer sequences for a number of murine, feline, and primate type C retroviruses. This alignment, in combination with genetic and biochemical studies, is proving useful for correlating sequence motifs with particular disease specificities (50) and has helped us to formulate a working model of enhancer design for this group of viruses.

The most highly conserved element in the enhancer of mammalian type C viruses is a sequence that includes the motif CAGGATA (Fig. 2A), recognized by several distinct nuclear factors but first defined as the binding site for a nuclear factor called LVb (64; Manley et al., unpublished). The LVb-binding site is invariably flanked on its 3' side by a somewhat less highly conserved TG(T/C)GGT (core) motif, which is frequently followed (5' → 3') by the NF1 and GRE consensus sequences. We view this assembly of highly conserved binding sites as the enhancer "framework."

Within the conserved motifs that constitute the enhancer framework, there are deviations from consensus sequences among viruses; at least some of these deviations may contribute to the distinct biological properties conferred by type C virus enhancers. Studies in three laboratories have implicated specific core motifs as conferring lymphoid-specific transcriptional activity to the MoMLV and SL3-3 enhancers

(5, 65, 72), and Clark et al. noted that most MLVs that induce erythroleukemias share a particular core motif (TGCGGTAAA) that is distinct from the core sequence from viruses that cause T cell lymphomas (15). Mutation of the core motif in MoMLV (TGTTGGAAA → TGCCGAAA) shifted the disease specificity from thymic leukemia to erythroleukemia (Speck et al., in press). Mutations in the adjoining LVb-binding site in MoMLV also resulted in a virus that induced some erythroleukemias. Together, these observations suggest that the conserved core motif and, possibly, LVb do play an important role in disease specificity.

Some viruses, however, have identical LVb/core sequences, yet genetic studies show that their enhancers encode different forms of leukemia. The nondefective Friend MLV clone 57 has a core sequence identical to that of MoMLV, yet the Friend MLV enhancer confers an erythroid rather than thymic disease specificity. This would argue that the highly variable sequences flanking the enhancer framework, particularly 5' and immediately 3' to the highly conserved LVb/core motif, can also specify disease phenotypes, and, indeed, genetic studies support this conclusion (27). Several nuclear factors have been identified that bind sites flanking the LVb/core motif in the enhancer regions of Friend MLV and MoMLV, and several of these sites are specific to either the Friend MLV or MoMLV enhancer (Fig. 4) (50, 64). The factors identified so far that can bind to the first copy of the direct repeats of MoMLV or Friend MLV are represented schematically in Fig. 4. It is apparent that the minor sequence differences between the two enhancers in regions flanking framework sequences can alter the array of factors that can assemble over the element. The factors that bind to variable regions flanking framework sequences might determine the distinct biological properties of the MoMLV and Friend MLV enhancers directly, or, alternatively, they might do so by influencing the association of cell-type-specific transcription factors to the more highly conserved LVb/core region.

We have used extremely high conservation of sequences among a set of closely related viruses to help us locate elements in the U3 region, particularly the enhancer, that are functionally important. Confirming the validity of this approach, we note that previously identified *cis*-acting elements, the IR, UCR, CCAAT, and TATA motifs, stood out in our analysis. Thus, we suspect that in addition to the enhancer sequences discussed above, at least two other highly conserved sequence elements we identified in this screen may be functionally significant. These motifs are located between CCAAT and TATA and may be important

promoter elements. Two additional 6-bp motifs, one between the direct repeat and the CCAAT box and one just 3' of the UCR site, were also seen (Fig. 1). We would, of course, expect that these identified motifs may be only a subset of the highly conserved binding sites for important proteins, since some sequence-specific DNA-binding proteins tolerate remarkably degenerate sequences (3, 59). It is also possible that some regions with only minor sequence variation might nevertheless bind discrete, independent factors in a manner analogous to that used by the putative core family of proteins discussed above.

A number of type C virus-related endogenous proviral LTR fragments have been cloned and sequenced, although to date their biological activity has been studied in only limited detail (8, 37, 41; J. Stoye, personal communication). To determine whether these viruses share the enhancer framework that we observed among exogenous type C viruses, we aligned 18 endogenous U3 sequences, including 6 belonging to the polytropic class of endogenous viruses and 4 of the modified polytropic group (data available upon request) (69). The UCR sequence, the TATA motif, and the two elements that were found between the CCAAT and TATA sequences in the exogenous viruses were highly conserved in the U3 regions of the endogenous viruses. A variation of the CCAAT sequence (to CCAAC) was more common among the endogenous viruses. In most cases, sequences were not available for the 5' edge of the LTR so the conservation of the IR sequence could not be determined. Within the enhancer region, there was considerable deviation from the highly conserved LVb/core motif of the exogenous viruses. CAGGA is deleted in the putative enhancers from all the polytropic viruses we examined, thus eliminating the LVb-binding site. Most of the endogenous viruses also lacked the highly conserved portion of the core motif [TG(T/C)GGT], having in its place the sequence GGTGGT. The NF1 site 3' of the LVb/core motif was frequently absent in the endogenous viruses through the removal of the 3' half of its consensus sequence (CGGN₅GCAA replaces CGGN₅₋₆GCCAA). The GRE site was retained in most viruses. Finally, as noted previously by others, many endogenous viruses have a 190-bp insert between their enhancer and promoter; this element may in itself provide an enhancer function (8, 36). Thus, the enhancer regions of these viruses were frequently distinct from those of the exogenous viruses.

Given the large numbers of endogenous LTR sequences in the mouse genome, it would be interesting to know more about the transcriptional activity of their distinctive LTRs. As with the exogenous viruses, it is the rich biology and extraordinary diversity of elements available that makes the study of transcriptional signals of mammalian type C viruses so interesting.

ACKNOWLEDGMENTS

We thank Will Gilbert, Director of the Whitehead Institute Computer Center, for expert help in setting up the retroviral data base and for instruction in the use of the University of Wisconsin Genetics Computer Group and NAQ programs. We thank John Coffin and Jonathon Stoye for particularly helpful discussions. We also thank Ellen Fanning, Mike Finney, and Karen Lech for thoughtful comments on the manuscript.

This work was supported by Public Health Service grant R01-CA19308 from the National Institutes of Health to N. Hopkins and partially by grant P01-CA42063 (Core) to P. A. Sharp. N. Speck was supported by a fellowship from the Medical Foundation.

LITERATURE CITED

- Abelson, H. T., and L. S. Rabstein. 1970. Influence of prednisolone on Moloney leukemogenic virus in BALB/c mice. *Cancer Res.* **30**:2208-2212.
- Abelson, H. T., and L. S. Rabstein. 1970. Lymphosarcoma: virus induced thymic independent disease in mice. *Cancer Res.* **30**:2213-2222.
- Baumruker, R., R. Sturm, and W. Herr. 1988. OBP100 binds remarkably degenerate octamer motifs through specific interactions with flanking sequences. *Genes Dev.* **2**:1400-1413.
- Bestwick, R. K., B. A. Boswell, and D. Kabat. 1984. Molecular cloning of biologically active Rauscher spleen focus-forming virus and the sequences of its *env* gene and long terminal repeat. *J. Virol.* **51**:695-705.
- Boral, A. L., S. A. Okenquist, and J. Lenz. 1989. Identification of the SL3-3 virus enhancer core as a T-lymphoma cell-specific element. *J. Virol.* **63**:76-84.
- Breathnach, R., and P. Chambon. 1981. Organization and expression of eukaryotic split genes coding for proteins. *Annu. Rev. Biochem.* **50**:349-383.
- Celander, D., B. L. Hsu, and W. A. Haseltine. 1988. Regulatory elements within the murine leukemia virus enhancer regions mediated glucocorticoid responsiveness. *J. Virol.* **62**:1314-1322.
- Ch'Ang, L.-Y., W. K. Yang, F. E. Myer, and D.-M. Yang. 1989. Negative regulatory element associated with potentially functional promoter and enhancer elements in the long terminal repeats of endogenous murine leukemia virus-related proviral sequences. *J. Virol.* **63**:2746-2757.
- Chatis, P. A., C. A. Holland, J. W. Hartley, W. P. Rowe, and N. Hopkins. 1983. Role for the 3' end of the genome in determining disease specificity of Friend and Moloney murine leukemia viruses. *Proc. Natl. Acad. Sci. USA* **80**:4408-4441.
- Chatis, P. A., C. A. Holland, J. E. Silver, T. N. Fredrickson, N. Hopkins, and J. W. Hartley. 1984. A 3' end fragment encompassing the transcriptional enhancers of nondefective Friend virus confers erythroleukemogenicity on Moloney murine leukemia virus. *J. Virol.* **52**:248-254.
- Chattopadhyay, S. K., B. M. Baroudy, K. L. Holmes, T. N. Frederickson, M. R. Lander, H. C. Morse III, and J. W. Hartley. 1989. Biologic and molecular genetic characteristics of a unique MCF virus that is highly leukemogenic in ecotropic virus-negative mice. *Virology* **168**:90-100.
- Chirigos, M. A., D. Scott, W. Turner, and K. Perk. 1968. Biological, pathological and physical characterization of a possible variant of a murine sarcoma virus (Moloney). *Int. J. Cancer* **3**:223-237.
- Clark, S. P., and T. W. Mak. 1982. Nucleotide sequences of the murine retrovirus Friend SFFV long terminal repeats: identification of a structure with extensive dyad symmetry 5' to the TATA box. *Nucleic Acids Res.* **10**:3315-3330.
- Clark, S. P., and T. W. Mak. 1984. Comparison of the sequences of the murine and gibbon ape retrovirus LTR's: analysis of elements involved in transcriptional control and provirus integration. *In* M. Pearson and N. Sternberg (ed.), *Gene transfer and cancer*. Raven Press, New York.
- Clark, S. P., R. Kaufhold, A. Chan, and T. W. Mak. 1985. Comparison of the transcriptional properties of the Friend and Moloney long terminal repeats: importance of tandem duplications and of the core enhancer sequence. *Virology* **141**:481-494.
- Corcoran, L. M., J. M. Addams, A. R. Dunn, and S. Cory. 1984. Murine T cell lymphomas in which the cellular *myc* oncogene has been activated by retroviral insertion. *Cell* **37**:113-122.
- DeFranco, D., and K. R. Yamamoto. 1986. Two different factors act separately or together to specify functionally distinct activities at a single transcriptional enhancer. *Mol. Cell. Biol.* **6**:993-1001.
- DesGroseillers, L., and P. Jolicœur. 1984. Mapping the viral sequences conferring leukemogenicity and disease specificity in Moloney and amphotropic murine leukemia viruses. *J. Virol.* **52**:448-456.
- DesGroseillers, L., E. Rassart, and P. Jolicœur. 1983. Thymotropism of murine leukemia virus is conferred by its long terminal repeat. *Proc. Natl. Acad. Sci. USA* **80**:4203-4207.

20. DesGroseillers, L., R. Villemur, and P. Jolicoeur. 1983. The high leukemic potential of Gross passage A murine leukemia virus is conferred by its long terminal repeat. *J. Virol.* **47**:24–32.
21. Devare, S. G., E. P. Reddy, J. D. Law, K. C. Robbins, and S. A. Aaronson. 1983. Nucleotide sequence of the simian sarcoma virus genome; demonstration that is acquired cellular sequences encode the transforming gene product p28sis. *Proc. Natl. Acad. Sci. USA* **80**:731–735.
22. Devereux, J., P. Haerberli, and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12**:387–395.
23. DeVries, E., W. van Driel, M. Tromp, J. H. van Bloom, and P. C. van der Vliet. 1985. Adenovirus DNA replication *in vitro*: site directed mutagenesis of the nuclear factor I binding site of the Ad2 origin. *Nucleic Acids Res.* **13**:4935–4952.
24. Dhar, R., W. L. McClements, L. W. Enquist, and G. F. Vande Woude. 1980. Nucleotide sequences of integrated Moloney sarcoma provirus long terminal repeats and their host and viral junctions. *Proc. Natl. Acad. Sci. USA* **77**:3937–3941.
25. Flanagan, J. R., A. M. Krieg, E. E. Max, and A. S. Khan. 1989. Negative control region at the 5' end of murine leukemia virus long terminal repeats. *Mol. Cell. Biol.* **9**:739–746.
26. Gilman, M. Z., R. N. Wilson, and R. Weinberg. 1986. Multiple protein-binding sites in the 5'-flanking region regulate *c-fos* expression. *Mol. Cell. Biol.* **6**:4305–4316.
27. Golemis, E., Y. Li, T. N. Fredrickson, J. W. Hartley, and N. Hopkins. 1989. Distinct segments within the enhancer region collaborate to specify the type of leukemia induced by nondefective Friend and Moloney viruses. *J. Virol.* **63**:328–337.
28. Graves, B. J., R. N. Eisenman, and S. L. McKnight. 1985. Delineation of transcriptional control signals within the Moloney murine sarcoma virus long terminal repeat. *Mol. Cell. Biol.* **5**:1948–1958.
29. Greenberg, M. E., Z. Zeigfried, and E. B. Ziff. 1987. Mutation of the *c-fos* gene dyad symmetry element inhibits serum inducibility of transcription *in vivo* and the nuclear regulatory factor binding *in vitro*. *Mol. Cell. Biol.* **7**:1217–1225.
30. Holland, C. A., C. Y. Thomas, S. K. Chattopadhyay, C. Koehne, and P. V. O'Donnell. 1989. Influence of enhancer sequences on thymotropism and leukemogenicity of mink cell focus-forming viruses. *J. Virol.* **63**:1284–1292.
31. Ishimoto, A., A. Adachi, K. Sakai, and M. Matsuyama. 1985. Long terminal repeat of Friend-MCF virus contains the sequence responsible for erythroid leukemia. *Virology* **141**:30–42.
32. Ishimoto, A., M. Takimoto, A. Adachi, M. Kakuyama, S. Kato, K. Kakimi, K. Fukuoka, T. Ogiu, and M. Matsuyama. 1987. Sequences responsible for erythroid and lymphoid leukemia in the long terminal repeats of Friend mink cell focus-forming and Moloney murine leukemia viruses. *J. Virol.* **61**:1861–1866.
33. Janowski, M., J. Merregaert, J. Boniver, and J. R. Maisin. 1985. Proviral genome of radiation leukemia virus: molecular cloning of biologically active proviral DNA and the nucleotide sequence of its long terminal repeat. *J. Virol.* **55**:251–255.
34. Johnson, P. F., W. H. Lanschulz, B. J. Graves, and S. L. McKnight. 1987. Identification of a rat liver nuclear protein that binds to the enhancer core element of three animal viruses. *Genes Dev.* **1**:133–146.
35. Jolicoeur, P., N. Nicolaiew, L. DesGroseillers, and E. Rassart. 1983. Molecular cloning of infectious viral DNA from ecotropic neurotropic wild mice retrovirus. *J. Virol.* **45**:1159–1163.
36. Kelly, M., C. A. Holland, M. L. Lung, S. K. Chattopadhyay, D. R. Lowy, and N. H. Hopkins. 1983. Nucleotide sequence of the 3' end of MCF247 murine leukemia viruses. *J. Virol.* **45**:291–298.
37. Khan, A. S., F. Laigret, and C. P. Rodi. 1987. Expression of mink cell focus-forming murine leukemia virus-related transcripts in AKR mice. *J. Virol.* **61**:876–882.
38. Khan, A. S., and M. Martin. 1983. Endogenous murine leukemia proviral long terminal repeats contain a unique 190-base-pair insert. *Proc. Natl. Acad. Sci. USA* **80**:2699–2703.
39. Kim, J. P., H. S. Kaplan, and K. E. Fry. 1982. Characterization of an infective molecular clone of the B-tropic, ecotropic BL/Ka(B) murine retrovirus genome. *J. Virol.* **44**:217–225.
40. Koch, W., W. Zimmerman, A. Oliff, and R. Friedrich. 1984. Molecular analysis of the envelope gene and long terminal repeat of Friend mink cell focus-forming virus: implications for the functions of these sequences. *J. Virol.* **49**:828–840.
41. Krieg, A. M., A. S. Khan, and A. D. Steinberg. 1988. Multiple endogenous xenotropic and mink cell focus-forming murine leukemia virus-related transcripts are induced by polyclonal immune activators. *J. Virol.* **62**:3545–3550.
42. Kriegler, M., and M. Botchan. 1983. Enhanced transformation by a simian virus 40 recombinant virus containing a Harvey murine sarcoma virus long terminal repeat. *Mol. Cell. Biol.* **3**:325–339.
43. Laimins, L. A., P. Gruss, R. Pozzatti, and G. Khoury. 1984. Characterization of enhancer elements in the long terminal repeat of Moloney murine sarcoma virus. *J. Virol.* **49**:183–189.
44. Leegwater, P. A. J., P. C. van der Vliet, F. A. W. Rupp, J. Nowock, and A. Suppel. 1986. Functional homology between the sequence-specific DNA-binding proteins nuclear factor 1 from HeLa cells and the TGGCA protein from chicken liver. *EMBO J.* **5**:381–386.
45. Lenz, J., D. Celander, R. L. Crowther, R. Patarca, D. W. Perkins, and W. A. Haseltine. 1984. Determination of the leukaemogenicity of a murine retrovirus by sequences within the long terminal repeat. *Nature (London)* **308**:467–470.
46. Levinson, B., G. Khoury, G. Vande Woude, and P. Gruss. 1982. Activation of SV40 genome by 72-base-pair tandem repeats of Moloney sarcoma virus. *Nature (London)* **295**:569–572.
47. Li, Y., E. Golemis, J. W. Hartley, and N. Hopkins. 1987. Disease specificity of nondefective Friend and Moloney murine leukemia viruses is controlled by a small number of nucleotides. *J. Virol.* **61**:693–700.
48. Liou, R. S., L. R. Boone, J. O. Kiggans, D. M. Yang, Y. W. Wang, R. W. Tennant, and W. K. Yang. 1983. Molecular cloning and analysis of the endogenous retrovirus chemically induced from RFM/Un mouse cell cultures. *J. Virol.* **46**:288–292.
49. Lung, M. L., J. W. Hartley, W. P. Rowe, and N. H. Hopkins. 1983. Large RNase T₁-resistant oligonucleotides encoding p15E and the U3 region of the long terminal repeat distinguish two biological classes of mink cell focus-forming type C viruses of inbred mice. *J. Virol.* **45**:275–290.
50. Manley, N. R., M. A. O'Connell, P. A. Sharp, and N. Hopkins. 1989. Nuclear factors that bind to the enhancer region of nondefective Friend murine leukemia virus. *J. Virol.* **63**:4210–4223.
51. Miksicsek, R., A. Heber, W. Schmid, U. Danesch, G. Posseckert, M. Beato, and G. Schutz. 1986. Glucocorticoid responsiveness of the transcriptional enhancer of Moloney murine sarcoma virus. *Cell* **46**:283–290.
52. Moloney, J. B. 1966. A virus-induced rhabdomyosarcoma of mice. *Natl. Cancer Inst. Monogr.* **22**:139–142.
53. Nagata, K., R. A. Guggenheimer, U. Inomoto, H. H. Lichy, and J. Hurwitz. 1982. Adenovirus replication *in vitro*: identification of a host factor that stimulates synthesis of the preterminal protein-dCMP complex. *Proc. Natl. Acad. Sci. USA* **79**:6438–6442.
54. Norton, J. D., J. Corror, and R. J. Avery. 1984. Genesis of Kirsten murine sarcoma virus: sequence analysis reveals recombination points and potential leukaemogenic determinant on parental leukaemia virus. *Nucleic Acids Res.* **12**:6839–6852.
55. O'Neill, R. R., C. E. Buckler, T. S. Theodore, M. A. Martin, and R. Repaske. 1985. Envelope and long terminal repeat sequences of a cloned infectious NZB xenotropic murine leukemia virus. *J. Virol.* **53**:100–106.
56. Ostrowski, M. C., D. Berard, and G. L. Hager. 1981. Specific transcriptional initiation *in vitro* on murine type-C retrovirus promoters. *Proc. Natl. Acad. Sci. USA* **78**:4485–4489.
57. Overhauser, J., and H. Fan. 1985. Generation of a glucocorticoid responsive Moloney murine leukemia virus by insertion of regulatory sequences from murine mammary tumor virus into the long terminal repeat. *J. Virol.* **54**:133–144.
58. Panganiban, A. T., and H. M. Temin. 1983. The terminal nucleotides of retrovirus DNA are required for integration but not virus production. *Nature (London)* **300**:155–160.

59. Pfeifer, K., T. Prezant, and L. Guarente. 1987. Yeast HAP1 activator binds to two upstream activation sites of different sequence. *Cell* **49**:19–27.
60. Prywes, R., and F. G. Roeder. 1986. Inducible binding of a factor to the *c-fos* enhancer. *Cell* **47**:777–784.
61. Reddy, E. P., J. J. Smith, and A. Srinivasan. 1983. Nucleotide sequence of Abelson murine leukemia virus genome: structural similarity of its transforming gene product to other onc gene products with tyrosine-specific kinase activity. *Proc. Natl. Acad. Sci. USA* **80**:3623–2627.
62. Shinnick, T. M., R. A. Lerner, and J. G. Sutcliffe. 1981. Nucleotide sequence of Moloney murine leukemia virus. *Nature (London)* **293**:543–548.
63. Sorge, J., D. Wright, V. E. Erdman, and A. E. Cutting. 1984. Amphotropic retrovirus vector system for human gene transfer. *Mol. Cell. Biol.* **4**:1730–1737.
64. Speck, N. A., and D. Baltimore. 1987. Six different nuclear factors interact with the 75-base-pair direct repeat of the Moloney murine leukemia virus enhancer. *Mol. Cell. Biol.* **7**:1101–1110.
65. Speck, N. A., B. Renjifo, and N. Hopkins. 1990. Point mutations in the Moloney murine leukemia virus enhancer identify a lymphoid-specific viral core motif and 1,3-phorbol myristate acetate-inducible element. *J. Virol.* **64**:543–550.
66. Stacey, A., C. Arbuthnot, R. Kollek, L. Coggins, and W. Ostertag. 1984. Comparison of myeloproliferative sarcoma virus with Moloney murine sarcoma virus variants by nucleotide sequencing and heteroduplex mapping. *J. Virol.* **50**:725–732.
67. Stewart, M. A., M. Warnock, A. Wheeler, N. Wilkie, J. I. Mullins, D. E. Onions, and J. C. Neil. 1986. Nucleotide sequences of a feline leukemia virus subgroup A envelope gene and long terminal repeat and evidence for the recombinational origin of subgroup B viruses. *J. Virol.* **58**:825–834.
68. Stocking, C., R. Killek, U. Bergholz, and W. Ostertag. 1986. Point mutations in the U3 region of the long terminal repeat of Moloney murine leukemia virus determine disease specificity of the myeloproliferative sarcoma virus. *Virology* **153**:145–149.
69. Stoye, J. D., and J. M. Coffin. 1987. The four classes of endogenous murine leukemia virus: structural relationships and potential for recombination. *J. Virol.* **61**:2659–2669.
70. Theodore, T. S., and A. S. Khan. 1987. Nucleotide sequence analysis of long terminal repeats of leukemogenic and nonleukemogenic MCF MuLVs. *Nucleic Acids Res.* **15**:5898.
71. Thiesen, H. J., Z. Bosze, L. Henry, and P. Charnay. 1988. A DNA element responsible for the different tissue specificities of Friend and Moloney retroviral enhancers. *J. Virol.* **62**:614–618.
72. Thornell, A., B. Hallberg, and T. Grundstrom. 1988. Differential protein binding in lymphocytes to a sequence in the enhancer of the mouse retrovirus SL3-3. *Mol. Cell. Biol.* **8**:1625–1637.
73. Trainor, C. D., J. L. Scott, S. F. Josephs, K. E. Fry, and M. S. Reitz, Jr. 1984. Nucleotide sequence of the large terminal repeat of two different strains of gibbon ape leukemia virus. *Virology* **137**:201–205.
74. Treisman, R. 1985. Transient accumulation of *c-fos* RNA following serum stimulation requires a conserved 5' element and *c-fos* sequences. *Cell* **42**:889–902.
75. Treisman, R. 1986. Identification of a protein-binding site that mediated transcriptional response of the *c-fos* gene to serum factors. *Cell* **46**:567–574.
76. Van Beveran, C., E. Rands, S. K. Chattopadhyay, D. E. Lowy, and I. M. Verma. 1982. Long terminal repeat of murine retroviral DNAs: sequence analysis, host-proviral junctions, and preintegration site. *J. Virol.* **41**:542–556.
77. Van Beveran, C., F. van Straaten, T. Curran, R. Muller, and I. M. Verma. 1983. Analysis of FBJ-MuSV provirus and *c-fos* (mouse) gene reveals that viral and cellular *fos* gene products have different carboxy termini. *Cell* **32**:1241–1255.
78. Van Beveran, C., F. van Straaten, J. A. Gallegher, and I. M. Verma. 1981. Nucleotide sequence of the genome of a murine sarcoma virus. *Cell* **27**:97–108.
79. Villemur, R., E. Rassart, L. DesGroseillers, and P. Jolicoeur. 1983. Molecular cloning of viral DNA from leukemogenic Gross passage A murine leukemia virus and nucleotide sequence of its long terminal repeat. *J. Virol.* **45**:539–546.
80. Vogt, M. 1982. Virus cloned from Rauscher virus complex induces erythroblastosis and thymic lymphoma. *Virology* **118**:226–236.
- 80a. Voytek, P., and C. A. Kozak. 1989. Nucleotide sequence and mode of transmission of the wild mouse ecotropic virus HoMuLV. *Virology* **173**:58–67.
81. Weiher, J., M. Zonig, and P. Gruss. 1983. Multiple point mutations affecting the simian virus 40 enhancer. *Science* **219**:626–631.
82. Weiss, R., N. Teich, H. Varmus, and J. Coffin (ed.). 1982. RNA tumor viruses, vol. 1. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
83. Weiss, R., N. Teich, H. Varmus, and J. Coffin (ed.). 1985. RNA tumor viruses, vol. 2. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
84. Yamamoto, K. R. 1985. Steroid receptor regulated transcription of specific genes and gene networks. *Annu. Rev. Genet.* **19**:209–215.