Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 3733–3737, May 1990
Genetics

Isolation of cellular promoters by using a retrovirus promoter trap

(transcriptional activation/insertional mutagenesis/polymerase chain reaction)

HARALD VON MELCHNER*, SITA REDDY, AND H. EARL RULEY

Center for Cancer Research and Department of Biology, Massachusetts Institute of Technology, 40 Ames Street, Cambridge, MA 02139

Communicated by Philip A. Sharp, February 21, 1990

ABSTRACT **A retrovirus vector has been used to isolate transcriptional promoters from mammalian cells. The virus contains a selectable gene encoding histidinol dehydrogenase (*his*) in the U3 region of the 3′ long terminal repeat (LTR). When the virus is passaged, duplication of LTRs places *his* sequences just 30 nucleotides from the adjacent cellular DNA. As a result, selection for histidinol resistance generates cell clones in which *his* is expressed on transcripts initiating in the flanking cellular DNA. Upstream cellular sequences, cloned after amplification by polymerase chain reaction, hybridized to RNA from uninfected cells, indicating that the adjacent promoters were transcriptionally active prior to virus integration. Two cloned transcribed flanking sequences also contained highly active transcriptional promoters, as estimated by their ability to activate expression of a linked reporter gene. Thus, U3His vectors provide a rapid and efficient means to isolate promoters active in different cell types. Moreover, by selecting for cell clones containing proviruses integrated in expressed genes, the virus may make an effective insertional mutagen.**

Transcription of eukaryotic genes is regulated by cis-acting DNA sequences. Promoters are located immediately upstream of transcriptional start sites and control transcriptional initiation by RNA polymerase. Enhancer elements increase the rate of transcriptional initiation and, to a certain extent, function irrespective of their position or orientation (1, 2).

Most promoters have been isolated from genomic libraries by using cDNA probes to identify sequences downstream of the transcriptional start site and by testing nearby sequences for promoter activity. However, isolating cellular promoters can be difficult because nearly full-length cDNA clones may be required to identify genomic sequences near the sites of transcriptional initiation, and transcribed genomic sequences may be hard to distinguish from untranscribed pseudogenes.

We have developed an alternative method to isolate transcriptional control elements which does not require prior isolation of gene sequences. The strategy exploited several features of the retroviral life cycle (3) to construct vectors that select for integration events near transcriptional promoters (4). In particular, retroviruses integrate at nearly random sites in the mammalian genome (5), yet recombination involves the precise recognition and conservation of specific viral sequences. Prior to integration, sequences near the ends of the viral genome duplicate such that the integrated provirus is flanked by long terminal repeat (LTR) sequences derived from the 5′ (U5) and 3′ (U3) ends of the viral genome. Foreign sequences can be inserted into the U3 region without adversely affecting the ability of the virus to be passaged; thus, histidinol-dehydrogenase-encoding sequences (*his*) inserted in U3 duplicated along with other LTR sequences. As a result, the *his* copy located in the 5′ LTR integrated just 30 nucleotides (nt) from the flanking cellular DNA. Selection for histidinol resistance generated cell clones in which the *his* gene in the 5′ LTR was invariably expressed on transcripts initiating nearby in the flanking cellular DNA (4). In this study, the polymerase chain reaction (PCR) (6) was used to isolate cellular promoters upstream of U3His proviruses.†

MATERIALS AND METHODS

Analysis of Proviral Flanking Sequences. Genomic DNAs from histidinol-resistant (His-ol^R) cell lines (4) were digested with *Hin*fI and ligated at a concentration of 5 μg/ml to obtain circular molecules. After cleavage with *Pvu* II, 1 μg of DNA from each sample was used for the PCR. PCRs were performed in 100 μl of 10 mM Tris·HCl, pH 8.3/5 mM KCl/1.5 mM $MgCl_2$/200 μM of each deoxyribonucleoside triphosphate, 2.5 units of Amplitaq DNA polymerase (Perkin–Elmer/Cetus), and each primer (5′-CCAGTCAATCAGGGTATTGA-3′ and 5′-GTCAGCGATATTCTGGATA-3′) at 1 μM. Reactions proceeded through 40 cycles of denaturation (95°C for 1.5 min), primer annealing (50°C for 1.5 min), and primer extension (72°C for 3 min). Gel-purified PCR products were cleaved with *Nhe* I and *Ssp* I and ligated to Bluescript KS(−) (Stratagene) plasmids digested with *Xba* I and *Eco*RV.

Nucleotide sequences of provirus–cell DNA junctions were determined by the dideoxy chain termination method as described (7).

Ribonuclease Protection Assays. Cellular RNA (30 μg) was hybridized to ^{32}P-labeled RNA probes as previously described (4). Probes complementary to the provirus flanking region coding strands extending from the U3 junction to a *Bss*HII restriction site 66 nt downstream of the junction were prepared by using phage T7 RNA polymerase (Promega Biotec) to transcribe the *Nhe* I/*Bss*HII fragments cloned in Bluescript KS(−). After hybridization, samples were digested with RNases A (Boehringer Mannheim) and T1 (BRL) at concentrations of 5 μg/ml and 1000 units/ml, respectively. Protected fragments were separated on denaturing 8% polyacrylamide/8.3 M urea gels and visualized by autoradiography.

Analysis of Promoter Activity. Provirus–cell DNA junctions (isolated after digesting pBluescript clones with *Not* I and *Hin*dIII) were ligated to pCAT (digested with *Bgl* II and *Hin*dIII). Prior to ligation, *Not* I and *Bgl* II ends were made blunt with the Klenow fragment of DNA polymerase. pCAT was derived from pTKCAT (8) by removing the thymidine kinase promoter sequences. Twenty micrograms of each pCAT plasmid together with 20 μg of pCH110 (Pharmacia) (9), a reporter plasmid expressing β-galactosidase, were cotransfected into NIH3T3 cells as previously described (10). After incubation for 48 hr, cells were recovered in 100 μl of 0.25 mM Tris·HCl, pH 8, and lysed by freeze-thawing. Chloramphenicol acetyltransferase (CAT) assays were performed as previ-

Abbreviations: CAT, chloramphenicol acetyltransferase; His-ol^R, histidinol-resistant; LTR, long terminal repeat; PCR, polymerase chain reaction; SV40, simian virus 40.

*To whom reprint requests should be addressed.

†The sequences reported in this paper have been deposited in the GenBank data base (accession nos. M33167 for P3aU3, M33168 for P3bU3, M33169 for P7U3, M33170 for Ψ7aU3, M33171 for Ψ7bU3, and M33172 for Ψ9U3).

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

ously described (11), using 1 mM chloramphenicol (Sigma) and 1 μCi of [^{3}H]acetyl-CoA (New England Nuclear, NET-290L, 200 mCi/mmol; 1 Ci = 37 GBq) per sample.

To control for variation in transfection efficiencies, CAT activity, expressed as cpm per milligram of cell protein [determined by the Bradford method (12)], was normalized for β-galactosidase expression as previously described (13).

RESULTS

Cloning of Cellular DNA Flanking the 5′ End of U3His Proviruses. The life cycle of the U3His vector is shown in Fig. 1. While the vector can transduce both neomycin and histidinol resistance, only 1 in 2500 U3His proviruses acquires the ability to express *his* at levels sufficient to render cells resistant to L-histidinol. Analysis of RNA extracted from His-ol^R cells suggested *his* expression resulted from transcripts initiating in the flanking cellular DNA (4).

Three out of four His-ol^R cell lines examined in the present study contained two proviruses, reflecting the initial multiplicity of infection (Table 1). To isolate cellular sequences that might have activated *his* expression, DNA flanking the 5′ end of U3His proviruses was amplified by the PCR.

Oligonucleotides were synthesized that would prime polymerase reactions in opposite directions from a *Pvu* II site located 72 nt from the end of the LTR (Fig. 2*A*). To generate small restriction fragments that amplify more efficiently, genomic DNA was digested to completion with *Hin*fI, yielding an average fragment length of 800 nt (data not shown). *Hin*fI fragments were circularized by using DNA ligase, thus positioning 5′ flanking sequences between the proviral priming sites. To avoid PCR products originating from circles formed at the 3′ end of the provirus, the DNA was digested with *Pvu* II, which cleaves fragments derived from the 3′ LTR (Fig. 2*A*). Such separation was less likely to occur at the left end because *Pvu* II sites are an order of magnitude less frequent than *Hin*fI sites in mammalian DNA.

PCR products from each His-ol^R line varied in size, as one might expect if flanking *Hin*fI sites were located at different distances from the proviruses. Accordingly, amplified fragments were 725–1130 nt in size, which corresponds to 95–500 nt of cellular DNA appended to a 630-nt U3*his* segment (Fig. 2 *B* and *C*; Table 1). In some clones, digestion by *Pvu* II was incomplete, resulting in a PCR product of 680 nt, derived from the 3′ end of the provirus (Fig. 2*B*). With the exception of Ψ9 cells, the number of amplification products matched the number of integrated proviruses (Table 1), indicating that in most instances it was possible to amplify upstream cellular sequences. However, this strategy is not expected to amplify upstream sequences in cases where the *Hin*fI sites are far apart or when the flanking fragment is cleaved by *Pvu* II.

Table 1. Cell lines and results obtained with cloned flanking sequences

His-ol^R cell line	Number of provirus integrations	Flanking DNA, nt	RNase protection	Promoter activity
P3	2	a 500	+	+
		b 180	−	−
P7	1	110	+	−
Ψ7	2	a 420	−	−
		b 300	+	+
Ψ9	2	a 95	+	−
		b NA		

The table lists the number of proviruses in each His-ol^R line (4) and sizes of flanking cellular sequences cloned after PCR. When there are two integrations, they are referred to as a and b, and the cloned fragments are named accordingly, e.g., P3a and P3b. Cloned flanking sequences that hybridized to transcripts from uninfected NIH3T3 cells and activated expression of a CAT reporter gene in a transient expression assay are indicated (+). NA, not amplified.

Amplified DNAs were digested with *Nhe* I/*Ssp* I and cloned in plasmid vectors digested with *Xba* I and *Eco*RV (Fig. 2 *A* and *C*). Sequence analysis of the cloned PCR products confirmed that each contained authentic junctions between viral and cellular DNA (Fig. 3). Each junction lacked the first 2 nt of U3 normally deleted during provirus integration (3); thus, *his* sequences did not interfere with recognition or ligation of sequences near the end of U3. U3 sequences were otherwise unaltered, except one provirus in Ψ7 cells contained a 4-nt substitution in the inverted repeat region (Fig. 3). This alteration was not a PCR artifact, since the same sequence was found in a fragment isolated from an independent amplification reaction. However, the mutation may have occurred after provirus integration, since similar mutations in U5 seriously affect viral replication (14).

The cellular portion of each sequence was unique (Fig. 3), indicating that each provirus was derived from an independent integration event. Computer analysis of the flanking sequences failed to identify similar sequences in the GenBank and EMBL data bases, except for the Ψ7b flanking sequence (Fig. 3) which was 84% identical to the consensus sequence of a B1 repetitive element (nt 89–38) (15, 16).

Sequences Flanking U3His Proviruses Are Complementary to Cellular Transcripts and Contain Active Promoters. In principle, provirus integration may affect the transcriptional activity of nearby cellular sequences, such that *his* expression results from insertional activation of cryptic cellular promoters. This seemed unlikely, since the U3His virus lacks known enhancers and only 1 in 2500 integrations was accompanied by *his* expression (4). Nevertheless, it was important

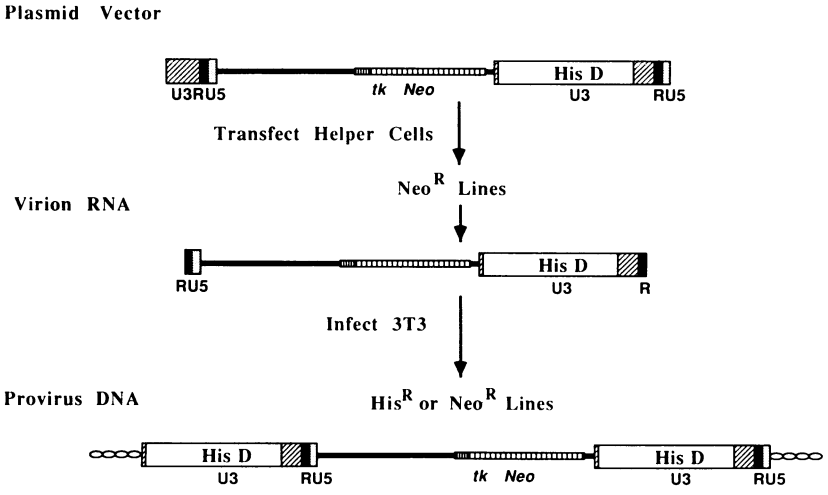

FIG. 1. Structure and replication of the U3His retrovirus vector. Gene sequences are labeled as follows: His D, *Salmonella typhimurium* histidinol dehydrogenase; *Neo*, neomycin phosphotransferase; and *tk*, herpes simplex virus thymidine kinase promoter; U3, R, and U5 are segments of the viral LTR. Infection of NIH3T3 cells generates proviruses flanked by *his* (see text). Neo^R, neomycin (G418)-resistant.

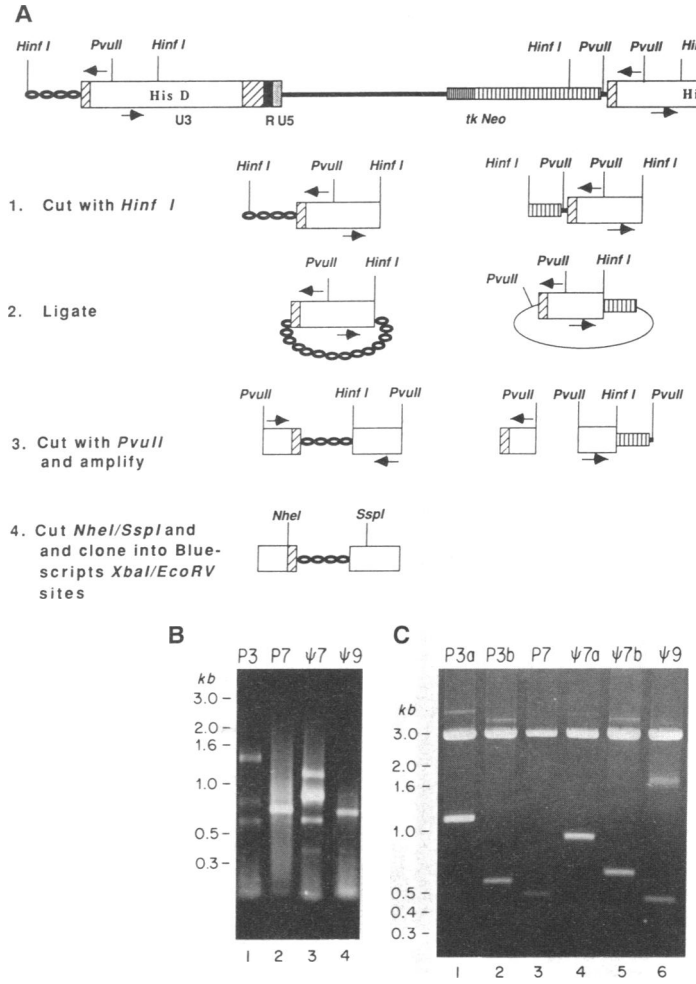

FIG. 2. Cloning of sequences flanking U3His proviruses. (*A*) Strategy for amplifying flanking sequences by PCR. (*B*) Ethidium bromide-stained gel showing the PCR products from His-ol^R cell lines P3, P7, Ψ7, and Ψ9. (*C*) Ethidium bromide-stained gel showing PCR products cloned in Bluescript vectors. kb, Kilobases.

to determine whether sequences which activate *his* expression are transcribed prior to provirus integration.

If flanking promoters were active prior to proviral integration, then sequences flanking the provirus should be expressed on cellular transcripts isolated from uninfected cells. ^{32}P-labeled RNA probes complementary to provirus-flanking sequences were prepared and hybridized to RNA extracted from 3T3 cells. Fig. 4*B* shows that RNA probes from each clone examined protected transcripts in NIH3T3 cells. However, such transcripts were detected by only one of the two probes isolated from cell lines with two proviruses, indicating that only one provirus had integrated into an expressed site. Protected fragments varied in size as one might expect from transcripts initiated in different genes (Fig. 4*B*). When hybridized to RNA from corresponding parental lines, two probes (P3a and Ψ7b) protected additional fragments that were 30 nt larger than the largest fragments protected by NIH3T3 RNA, exactly the size expected for transcripts initiated in the allele containing the provirus (Fig. 4).

Although flanking probes P7 and Ψ9 hybridized to RNA from NIH3T3 cells, transcripts containing virus-cell junctions were not detected in P7 and Ψ9 cells. The reasons for this are not known. Short probes complementary to LTR *his* sequences failed to detect cryptic splice sites that could have deleted 5′ U3 sequences. Instead, RNA from all His-ol^R clones examined protected fragments of the exact size expected for cellular transcripts extending through the 5′ end of the provirus (data not shown). It is possible that mutations introduced during PCR prevented allelic transcripts from being detected, since *Taq* DNA polymerase introduces mutations at a fairly high frequency (6) and even single nucleotide substitutions increase the sensitivity of RNA·RNA hybrids to RNases (17).

Some fragments protected by the Ψ7b probe probably result from unrelated transcripts hybridizing to the 94-nt B1

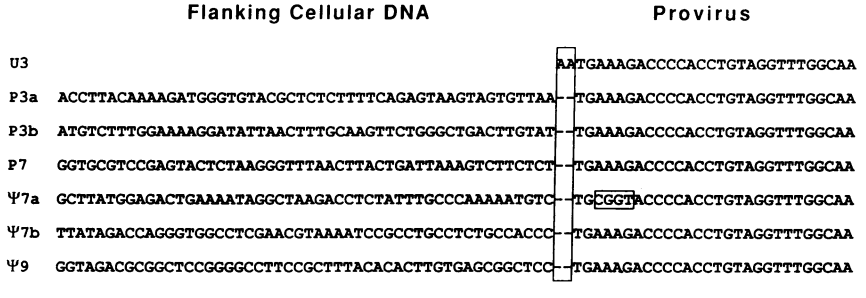

FIG. 3. Nucleotide sequences of the junctions between U3His provirus and flanking cellular DNA. Junctions and nucleotide substitutions in U3 are boxed; - indicates U3 nucleotide not present.

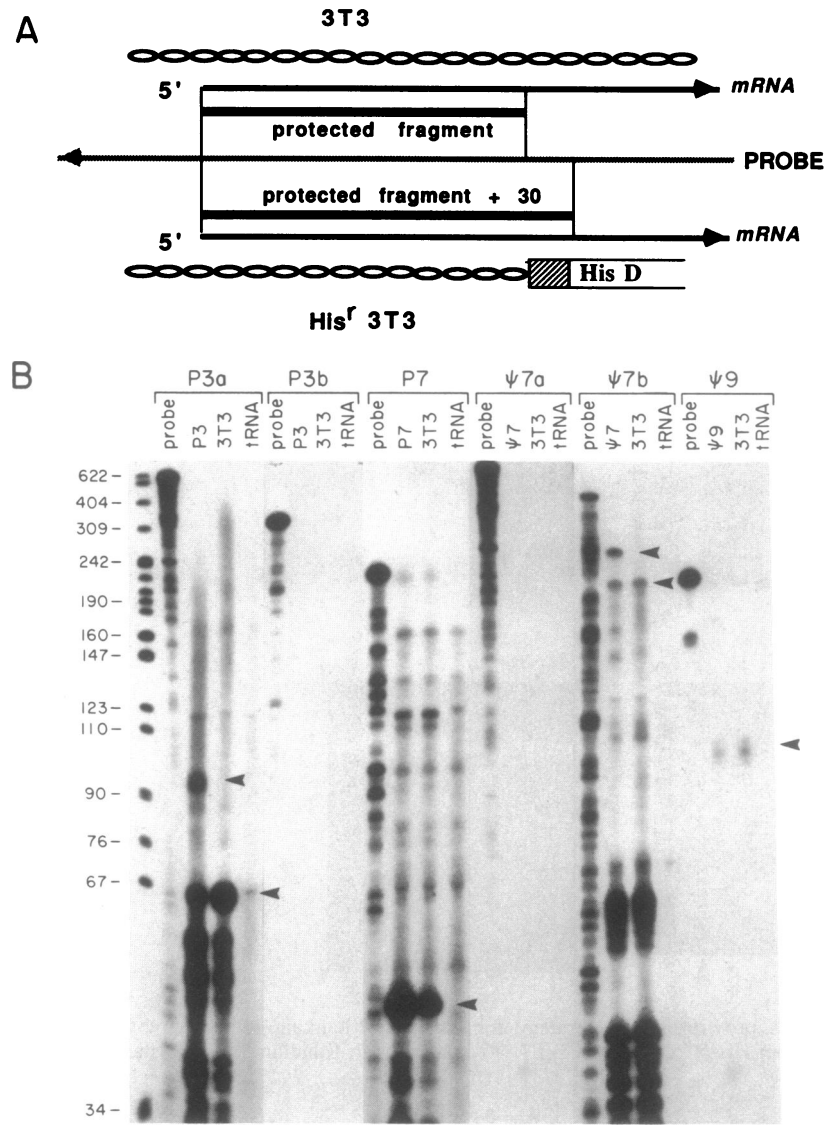

FIG. 4. Ribonuclease protection analysis of cellular transcripts. ^{32}P-labeled probes complementary to provirus-flanking sequence coding strands were prepared and hybridized each to RNA from NIH3T3 and parental cells. (*A*) Protected fragments expected in NIH3T3 cells and parental lines (His-ol^R 3T3) are indicated by solid bars. (*B*) RNA samples from NIH3T3 cells, parental His-ol^R lines, and tRNA were hybridized to P3a, P3b, P7, Ψ7a, Ψ7b, and Ψ9 probes, as indicated by brackets. Sizes of end-labeled pBR322 *Msp* I restriction fragments are indicated at the left; protected fragments are designated by arrowheads at the right.

sequence flanking the provirus, since B1 repeats are present in cellular mRNAs (15, 16). However, the orientation of B1 sequences is such that the Ψ7b probe should not detect normal B1 products transcribed by RNA polymerase III.

Finally, to investigate the ability of transcribed flanking sequences to activate expression of a linked reporter gene, provirus–cell DNA junctions were cloned in pCAT expression vectors. When tested for their ability to activate the expression of the CAT gene in transient transfection assays, two out of four transcribed flanking sequences stimulated CAT expression at a level greater than the simian virus 40 (SV40) early region promoter (Fig. 5, Table 1).

DISCUSSION

A retrovirus vector (U3His) has been used to isolate transcriptional promoters from mammalian cells. U3His vectors replicate normally when passaged to recipient cells and generate proviruses flanked by *his*. By selecting for *his* expression, cell clones were obtained in which the *his* gene in the 5′ LTR was expressed on transcripts initiating in the nearby cellular DNA (4).

A unidirectional PCR method has been developed which greatly facilitates cloning sequences immediately upstream of U3His proviruses. While a similar method has been described to clone proviral flanking sequences (18), U3His vectors simplify the procedure, since primers used for the PCR reaction are complementary to *his* rather than the viral LTR and thus do not prime sequences derived from endogenous retroviruses. Cloned flanking sequences hybridized to RNA from uninfected cells, and they also contained functional promoters. Thus U3His viruses select for instances in which the provirus has usurped active cellular promoters.

A number of functional assays to identify and clone mammalian transcriptional control elements have been described (19–23). Each involves (*i*) introducing gene coding sequences into cells, either by DNA-mediated gene transfer or by retrovirus transduction, (*ii*) selecting for cellular phenotypes that result when the gene is expressed, and (*iii*) analyzing linked DNA sequences for transcriptional activity. However, these strategies suffer from a variety of limitations. DNA-mediated gene transfer is relatively inefficient, and recombination between the selectable marker and transcriptional activator is unpredictable. As a result, the structure and mechanism of activation of the integrated gene vary from clone to clone. While retroviruses integrate by precise recombinational events, sequences in the 5′ LTR—in particular the viral enhancer, polyadenylylation signal, and multiple initiation and termination codons—interfere with transcriptional activation or translation of proviral genes. Vectors in which the retroviral enhancer has been deleted or which rely on RNA splicing to remove 5′ viral sequences have been used to identify transcriptionally active DNA (22, 24). However, activating cellular sequences are not closely linked to the provirus and are therefore not easily cloned.

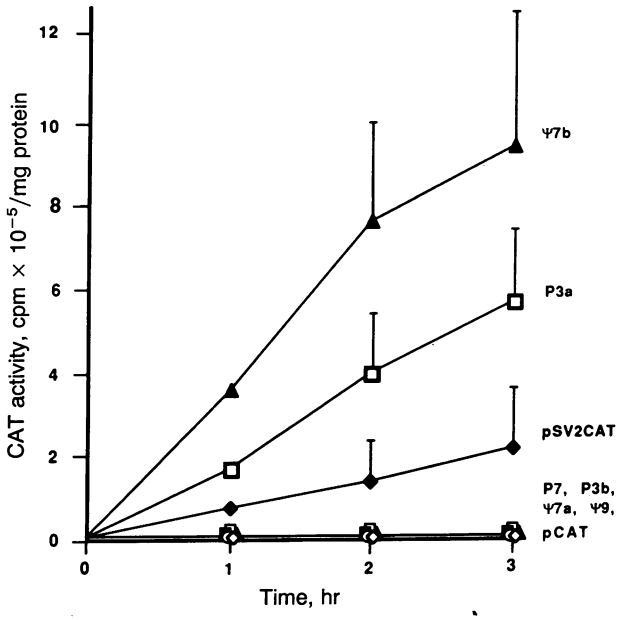

FIG. 5. Analysis of provirus flanking sequences for promoter activity. Flanking sequences cloned upstream of a CAT reporter gene were transfected into NIH3T3 cells and CAT expression was monitored. Results are normalized for transfection efficiency. Means and SEMs from four independent experiments are plotted. Flanking sequence designations are as for Fig. 2; pCAT (negative control) contains only the CAT coding sequence; pSV2CAT shows CAT expression from the SV40 early promoter.

The U3His vector avoids these problems because duplication of the LTR places *his* coding sequences just 30 nt from the flanking cellular DNA, outside of viral sequences that could interfere with transcriptional activation by upstream cellular promoters. Furthermore, several observations suggest that proviruses which transduce histidinol resistance integrate within a few hundred nt of the transcriptional start site. First, less than 200 nt of cellular sequence is appended to *his* coding sequences, as determined by Northern blot hybridization (4). Second, the maximum length of cellular RNA protected by flanking sequence probes was less than 250 nt. Finally, PCR products of 300 and 500 nt each contained functional promoters (Table 1).

The sequence of the 5′ end of the U3His vector provides a likely explanation for the close physical association between the activating cellular promoter and the integrated *his* gene. An amber (TAG) termination codon is located 24 nt upstream and in frame with the initiating AUG codon for histidinol dehydrogenase (data not shown). As a result, histidinol dehydrogenase cannot be synthesized as a fusion protein. This imposes a strong selection for integration events in which *his* provides the first initiating AUG in the resulting hybrid transcript. For the average mammalian gene, the provirus must integrate within 100 nt of the transcriptional start site (25).

Two upstream sequences contained functional promoters, both of which were more active than the SV40 early region promoter. This result was not unexpected, since steady-state levels of hybrid *his* transcripts were quite abundant as judged by Northern blot hybridization (4). Moreover, histidinol is thought to kill mammalian cells by acting as a competitive inhibitor of histidine–tRNA aminoacyltransferase (26) and, as a result, relatively high levels of histidinol dehydrogenase may be required to confer resistance. These considerations suggest that U3His viruses select for integrations near relatively strong promoters. However, it should be possible to isolate weaker promoters by using other selectable markers or less stringent selection strategies.

Although only 1 in 2500 integrations places a provirus in an appropriate position downstream from a promoter strong enough to confer histidinol resistance, it is still relatively easy to obtain 10^4 independent His-ol^R clones, the maximum estimated number of integration sites capable of activating *his* (4). Isolation of upstream sequences does not require extensive subculturing of His-ol^R clones (or pools of clones), since sufficient DNA for PCR amplification can be isolated from 5 × 10^5 cells (27). Thus, it should be feasible to screen a large number of flanking sequences for promoter activity after inserting PCR products directly into an expression vector.

In conclusion, U3His vectors should provide a rapid and efficient means to isolate a large number of transcriptional promoters active in different cell types. Transcribed flanking sequences may be used to isolate cDNAs derived from genes disrupted by virus integration or to map transcriptionally active sites to distinct chromosomal regions. Finally, U3His vectors may make effective insertional mutagens, since they select for instances in which integration occurs into expressed genes.

We thank Dr. Wen Chang for her help in performing the CAT assays. This work was supported by a grant from the National Cancer Institute (PO1 CA42063) to H.E.R. and partially supported by a Cancer Center Core Grant (CA14051). H.v.M. is a Heisenberg Fellow of the Deutsche Forschungsgemeinschaft (Bonn-Bad Godesberg, West Germany). H.E.R. is a Latham Professor of Biology.

1. Dynan, S. (1989) *Cell* **58,** 1–4.
2. Maniatis, T., Goodbourn, S. & Fischer, J. A. (1987) *Science* **236,** 1237–1245.
3. Varmus, H. (1988) *Science* **240,** 1427–1435.
4. von Melchner, H. & Ruley, H. E. (1989) *J. Virol.* **63,** 3227–3233.
5. Shih, C. C., Stoye, J. P. & Coffin, J. M. (1988) *Cell* **53,** 531–537.
6. Erlich, H. A., ed. (1989) *PCR Technology* (Stockton, New York).
7. Reddy, S., Ozgur, K., Lu, M., Chang, W., Mohan, S. R., Kumar, C. C. & Ruley, H. E. (1990) *J. Biol. Chem.* **265,** 1683–1687.
8. Burch, J. B. E., Evans, M. I., Friedman, T. M. & O'Malley, P. J. (1988) *Mol. Cell. Biol.* **8,** 1123–1131.
9. Hall, C., Jacob, E., Ringold, G. & Lee, F. (1983) *J. Mol. Appl. Genet.* **2,** 101–109.
10. Luthman, H. & Magnusson, G. (1983) *Nucleic Acids Res.* **11,** 1295–1307.
11. Neumann, J. R., Morency, C. A. & Russian, K. O. (1987) *BioTechniques* **5,** 444–447.
12. Bradford, M. (1976) *Anal. Biochem.* **72,** 248–254.
13. Guarente, L. & Ptashne, M. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 2199–2203.
14. Colicelli, J. & Goff, S. P. (1985) *Cell* **42,** 573–580.
15. Georgiev, G. P. (1984) *Eur. J. Biochem.* **145,** 203–220.
16. Weiner, A. M., Deininger, P. L. & Efstratiadis, A. (1986) *Annu. Rev. Biochem.* **55,** 631–661.
17. Myers, R. M., Larin, Z. & Maniatis, T. (1985) *Science* **230,** 1242–1246.
18. Silver, J. & Keerikatte, V. (1989) *J. Virol.* **63,** 1924–1928.
19. Fried, M., Griffith, M., Davies, B., Bjursell, G., Mantia, G. L. & Lania, L. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 2117–2121.
20. Weber, F., de Villiers, J. & Schaffner, W. (1984) *Cell* **36,** 983–992.
21. Bhat, K., McBurney, M. & Hamada, H. (1988) *Mol. Cell. Biol.* **8,** 3251–3259.
22. Brenner, D. G., Lin-Chao, S. & Cohen, S. N. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 5517–5521.
23. Gossler, A., Joyner, A. L., Rossant, J. & Skarnes, W. C. (1989) *Science* **244,** 463–465.
24. Barklis, E., Mulligan, R. C. & Jaenisch, R. (1986) *Cell* **47,** 391–399.
25. Kozak, M. (1987) *Nucleic Acids Res.* **15,** 8125–8148.
26. Hartman, S. C. & Mulligan, R. C. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 8047–8051.
27. Maniatis, T., Fritsch, F. & Sambrook, J. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), p. 9.19.