

# Retroviral characteristics of the long terminal repeat of murine E.Tn sequences

Mourad Kaghad, Laurence Maillet and Philippe Brûlet

Unité de Génétique cellulaire du Collège de France et de l'Institut Pasteur, 25, rue du Dr. Roux, 75724 Paris Cedex 15, France

Communicated by F.Jacob

**E.Tn sequences form a family of long moderately repeated sequences which are abundantly transcribed in the pluripotent cell lineage between day 3.5 and 7.5 of early mouse embryogenesis. The structure of the long terminal repeat (LTR) bordering the E.Tn sequences has been investigated by nucleotide sequencing, primer extension and S1 mapping experiments. This has allowed the identification of U3, R and U5 domains, and of several other structural features all of which are characteristics of retroviral LTRs.**

*Key words:* early mouse embryogenesis/endogenous retrovirus/early transposon

## Introduction

E.Tn sequences form a family of long moderately repeated sequences which are dispersed throughout the mouse genome. This family has been identified during studies on molecular aspects of the mouse blastocyst formation. The distinctive feature of this family is its transcription at a high level in undifferentiated embryonal carcinoma (EC) cell lines but at low or undetectable levels in the differentiated cell lines that we have tested. *In situ* hybridization on embryos reveals E.Tn RNA in great amounts in the pluripotent cells of the inner cell mass and embryonic ectoderm. From a genomic DNA library, an E.Tn sequence was isolated and subcloned into the pMAC-2 plasmid. A direct repeat, several hundred base pairs long, borders this E.Tn sequence. By R-loop analysis the E.Tn RNA was shown to be co-linear with the E.Tn sequence in the pMAC-2 plasmid (Brûlet *et al.*, 1983, 1985).

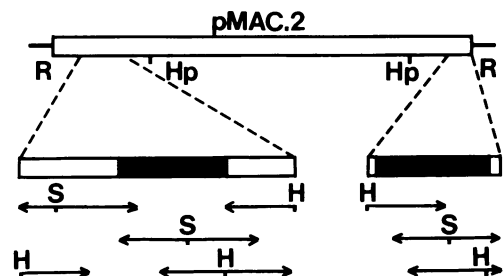
Here we report a detailed analysis of the two direct long terminal repeats (LTRs) that border the E.Tn sequence in pMAC-2. By using nucleotide sequencing and by applying primer extension and S1 mapping techniques, we have located the 5' and 3' ends of the RNA within the two direct LTRs of the E.Tn sequence. The structure of these two LTRs is essentially identical to the U3, R, U5 structure of retroviral LTRs while their nucleotide sequence differs from known LTR sequences.

## Results

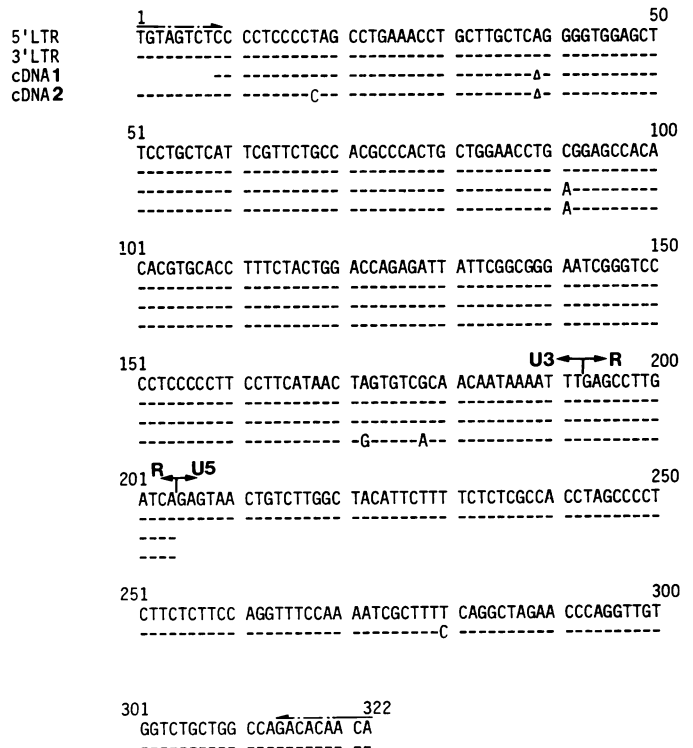
### Nucleotide sequence analysis of the LTR

Electron microscopy of heteroduplexes between the 6-kb DNA insert in pMAC-2 and the RNA transcript in EC cells showed that the 5' and 3' ends of the RNA are localized at the extremities of the insert (Brûlet *et al.*, 1983). The ends of the insert were electrophoretically purified from an *EcoRI*, *HpaII* digest of pMAC-2. These fragments were further digested with other restriction enzymes and subcloned into M13mp8 and M13mp9. The DNA was then sequenced by the dideoxy chain termination method (Sanger *et al.*, 1977; Messing and Vieira, 1982). The boundaries of the LTRs were determined by the divergence in

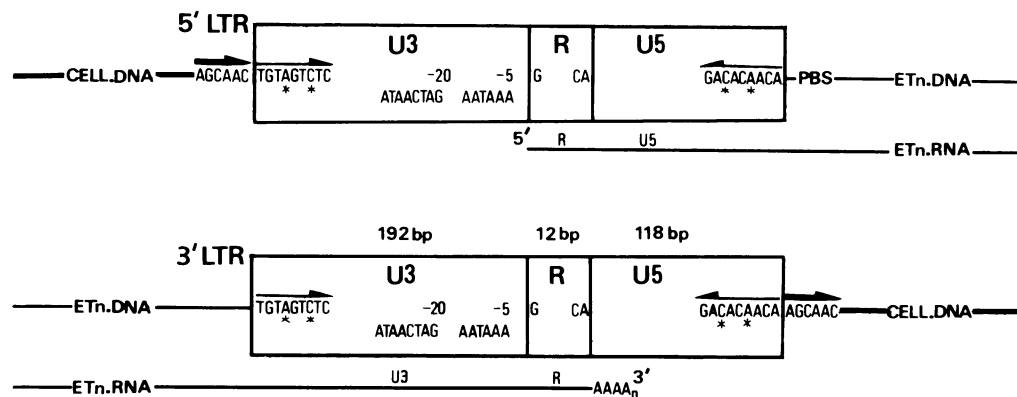
the nucleotide sequence between the left and right LTRs. Figure 1 illustrates our sequencing strategy. Figure 2 shows the nucleotide sequences of both LTRs. They are 322 bp long and differ by a single base mutation, a T exchanged with a C at position 280. An inverted repeat, nine nucleotides long with two mismatched nucleotides, borders the two LTRs. A 6-bp direct repeat, AGCAAC, brackets the entire E.Tn DNA sequence (Figure 3).



**Fig. 1.** Nucleotide sequencing strategy of E.Tn LTRs. The 1.2- and 0.8-kb *EcoRI*, *HpaII* fragments from the 6-kb genomic DNA insert in pMAC-2 were purified and, after further digestion, subcloned into M13mp8 (see Materials and methods). Arrows represent the extent of sequencing around the restriction sites. H: *HaeIII*; S: *Sau3A*; R: *EcoRI*; Hp: *HpaII*. U3, R and U5 are defined in the text.



**Fig. 2.** The nucleotide sequence of the 5' and 3' LTRs from the genomic E.Tn sequence in pMAC-2 and of two E.Tn cDNAs. Positions are numbered from the 5' end of the LTR sequence. A dashed line indicates identical sequence, Δ a deletion



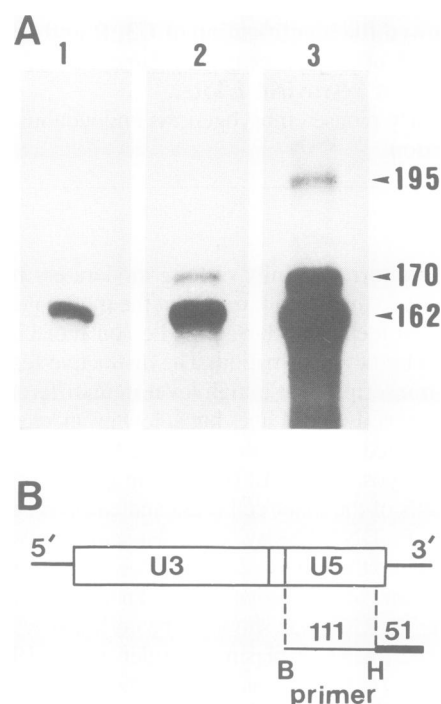
**Fig. 3.** Diagram of E.Tn LTRs. U3, R and U5 are the regions discussed in the text and are respectively 192, 12 and 118 bp long. Opposed arrows at the end of the LTR: a 9-bp inverted repeat with two mismatches indicated by asterisks. The 6-bp direct repeat in the immediate flanking genomic DNA is indicated by heavy arrows. PBS is a sequence homologous to tRNA<sub>3<sup>Lys</sup></sub>. Distances of the TATA box and the polyadenylation signal are from the U3/R boundary. The 5' and 3' ends of the RNA transcript are indicated.

The nucleotide sequence of the genomic LTR was also compared with the sequences of the two cDNAs obtained by reverse transcription of an oligo(dT)-primed RNA, and corresponding to the 3' end of the E.Tn RNA. These two cDNAs were obtained by screening, with an LTR-specific probe, a cDNA library constructed with  $\lambda$ gt11 and poly(A) RNA extracted from an EC cell line. The two DNA sequences are nearly identical to the LTR sequence up to nucleotide 204, where the cDNAs terminate with a poly(A) tail (Figure 2). Therefore, we localize the 3' end of E.Tn RNA at nucleotide 204 in the 3' LTR. The cDNA-1 has one substitution and one deleted nucleotide. The two DNAs differ by three mismatches implying that several E.Tn sequences are transcribed in EC cell lines.

#### Mapping the RNA initiation sites by primer extension

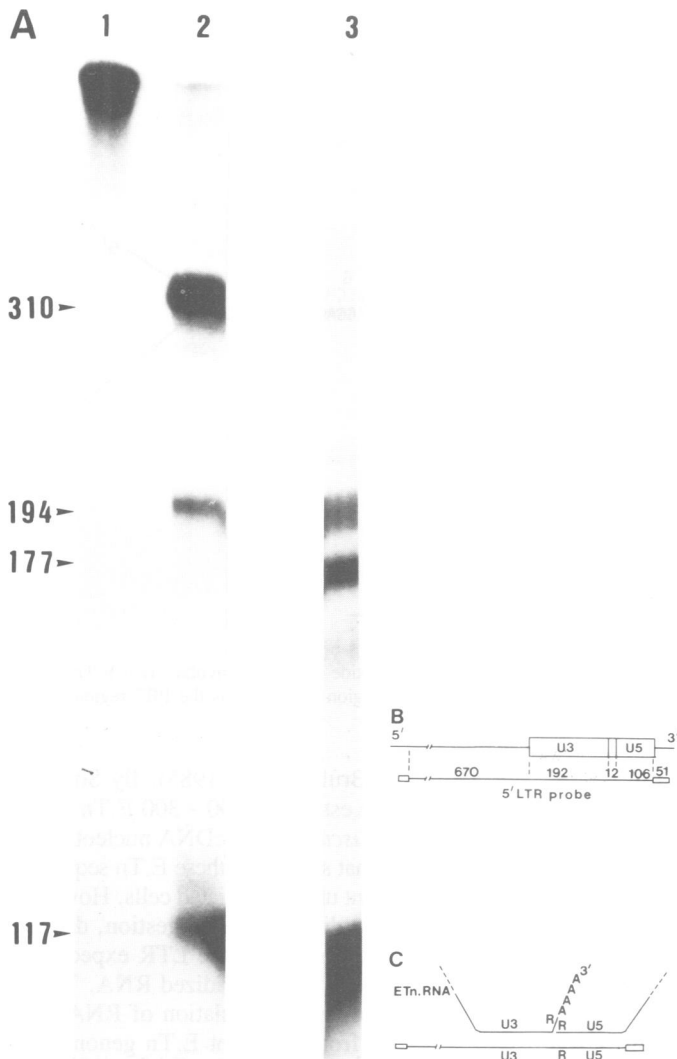
The primer extension method was used to localize the 5' end of the RNA. For this purpose, uniformly labelled single-stranded DNA molecules from the 5' end of the E.Tn sequence in pMAC-2 were synthesized (Figure 4). After hybridization to RNA from undifferentiated EC cells we checked that the radioactive fragment used for primer extension was entirely protected from S1 nuclease digestion (data not shown). The RNA-primer hybrid molecules were then elongated by reverse transcription. After denaturation, the length of the fragments was measured on sequencing gels using the four sequencing reaction ladders of the LTR fragments as size markers. As seen in Figure 4, the primer molecule, whose 3' end is at position 200 (see Figure 2), is elongated by eight bases. Therefore, the 5' end of the RNA is approximately at position 192. The nearest G is at position 193 and could be the initiation site of the RNA by analogy with retroviral RNA. Because we know from the cDNA sequencing data (see above) that nucleotides 193–204 are also at the 3' end of the RNA, this result identifies a 12 nucleotide long sequence which is duplicated at the 3' and 5' ends of the RNA. By analogy with retroviral RNAs and LTRs (Temin, 1981; Varmus, 1982), an R domain in the E.Tn LTR can thus be defined starting with a G at position 193 and ending with CA at position 203–204. The first 192 nucleotides of the LTR sequence, which we found at the 3' end of the E.Tn RNA prior to the R sequence, will hereafter be referred to as the U3 region. Similarly, the domain extending from nucleotide 205 to 322, which is found at the 5' end of the RNA after the R sequence, is referred to as the U5 region (Figures 2 and 3).

Twenty base pairs 5' from the postulated RNA initiation site



**Fig. 4.** Primer extension analysis of E.Tn RNA from EC cells. (A) The primer was hybridized to 2  $\mu$ g of poly(A) RNA for 5 h and extended with reverse transcriptase as described in Materials and methods. The products were analyzed on 8% polyacrylamide gels. The size of the fragments was estimated from sequencing reaction ladders. **Lanes 1 and 2:** primer before and after elongation. Autoradiography was for 2 h at  $-20^{\circ}\text{C}$ . **Lane 3:** same as 2 but autoradiography was overnight. (B) The single-stranded uniformly labelled primer is composed from its 5' end of 51 nucleotides from the M13mp8 and 111 nucleotides from the HaeIII site (H) at position 310 to the BclI site (B) at position 200 in the LTR (Figure 2).

is a sequence ATAACTAG which could serve as a Goldberg-Hogness TATA box. Also 5 bp 5' from the same initiation site is a probable adenylation signal AATAAA (Figure 3). Longer autoradiography reveals fainter bands (Figure 4). They might represent minor RNAs initiated from weaker promoters in the LTR (Allan *et al.*, 1983) or alternatively reflect a heterogeneity in the RNA population as different E.Tn genomic sequences were transcribed.



**Fig. 5.** (A) S1 nuclease analysis of E.Tn RNA from EC cells. Single-stranded 5'-LTR probe was hybridized to 1  $\mu$ g of poly(A) RNA for 5 h and digested with S1 nuclease (Brület *et al.*, 1985). The products of the digestion were analyzed on 8% polyacrylamide gels. Autoradiography was for 2 h at  $-80^{\circ}\text{C}$ . Lane 1 gives the position of the probe. The S1 digestion was performed for 2 h at 20 and  $50^{\circ}\text{C}$ , lane 2 and 3, respectively. (B) From its 5' end the probe is composed of 51 nucleotides from the vector M13mp8, the LTR from the *Hae*III site at position 310, 670 nucleotides from mouse genomic flanking region and 28 bases from the M13mp8 polylinker (*Eco*RI site). (C) Postulated RNA-DNA structure. The 5' end and 3' end of the E.Tn RNA protect, respectively, the R.U5 and U3 domains of the probe.

#### S1 mapping of the 5' and 3' ends of E.Tn RNA

A single-stranded 5' LTR probe obtained from an M13mp8 recombinant molecule comprises 670 bases upstream of the LTR and the entire LTR except for the last 12 bases of its 3' end (Figure 5B). After hybridization to RNA, the hybrids were digested with S1 nuclease under various stringency conditions. The denatured protected fragments were fractionated on sequencing gels. The results of this analysis are shown in Figure 5A. When the S1 digestion was performed at  $20^{\circ}\text{C}$ , three DNA fragments were obtained. Their sizes are estimated to be  $\sim 310$ , 194 and 117 bases and correspond, respectively, to the sizes of the LTR, U3 and R.U5 domains of the probe. From the primer extension experiment and from the sequence data (see above), we know that R.U5 is complementary to the RNA 5' end, while U3.R

corresponds to the 3' end. A looped RNA structure could protect the full length LTR, the 5' end of the RNA protecting R.U5 and the 3' end protecting U3 (Figure 5C), generating the protected DNA fragment of 310 bases. The two other protected fragments of 194 and 117 bases, which correspond respectively to the sizes of the U3 and R.U5 domains of the probe, could result from partial digestion at the junction U3/R.U5 predicted by the postulated looped RNA structure (Figure 5C). This result corroborates the overall structure of the LTR given in Figure 3 and obtained by nucleotide sequencing and primer extension experiments.

When the S1 digestion was performed at higher stringency,  $50^{\circ}\text{C}$  or  $37^{\circ}\text{C}$  for longer time, the 310-base protected DNA fragment disappears corresponding to a complete digestion at the U3/R.U5 junction. The band of 117 bases corresponding to R.U5 is not affected. However, the band of 194 bases ascribed to the U3 part of the LTR is split into several bands (Figure 5A). Several protected fragments are also obtained when a cDNA probe spanning U3 and R is used in the S1 digestion (P. Blanchet and P. Brület, unpublished results). Most probably the fragments arise from the digestion of S1-sensitive regions in U3. For instance, a homopyrimidine track, nucleotides 6–18, or an AT-rich domain, nucleotides 180–190, could be digested under stringent conditions by the S1 nuclease. The heterogeneity of the E.Tn RNA could also contribute to this pattern.

#### Inner junctions of the 5' and 3' LTR

Two base pairs 3' from the 5'-LTR there is a sequence of 14 out of 16 bases which is complementary to the 3' end of the tRNA<sup>Lys</sup> (Weiss *et al.*, 1982). This putative primer binding site (PBS) is followed by a stretch of 193 nucleotides (corresponding to the so-called region L in retroviruses), without an open reading frame. We find in the L region of E.Tn, between bases 402 and 410, a possible splice donor site. Its sequence 5'AGG-ATAAGG3' matches the consensus sequence 5'AGGTAAGT3' (Seif *et al.*, 1979) except for one A insertion and one substitution from T to G. Finally, we note that a possible dimer-linkage structure can be formed with two E.Tn RNA molecules and two tRNA<sub>3</sub><sup>Lys</sup> molecules (Figure 6). With the numbering used for the LTR in Figure 2, this structure would involve nucleotides 253–372, i.e., the 3' part of U5, the primer binding site PBS and part of the L region. The inner junction of the 3' LTR is composed of a 13-base sequence of A and G, and 3' to it 24 out of 26 bases are A and T.

#### Discussion

The structural analysis presented here establishes that the two LTRs bordering the long moderately repeated E.Tn sequences are similar to retroviral LTRs. The E.Tn LTR can be subdivided into three domains U3, R and U5, respectively 192, 12 and 118 bp long. The central R sequence of 12 bases is duplicated at both ends of E.Tn RNA while the U3 and U5 sequences are found at the 3' and 5' ends of the RNA. A previous electron microscopic analysis of heteroduplexes had shown that E.Tn RNA is co-linear with the randomly selected genomic E.Tn sequence of plasmid pMAC-2. The E.Tn LTR structure, as elucidated here, adds support to the notion that E.Tn sequences are organized in a way similar to retroviral sequences (Weiss *et al.*, 1982).

This conclusion is also supported by several details of the E.Tn LTR sequence. Twenty bases pairs 5' from the RNA initiation site is a putative TATA box sequence and a polyadenylation signal is located 5 bp 5' from the initiation site. These two consensus

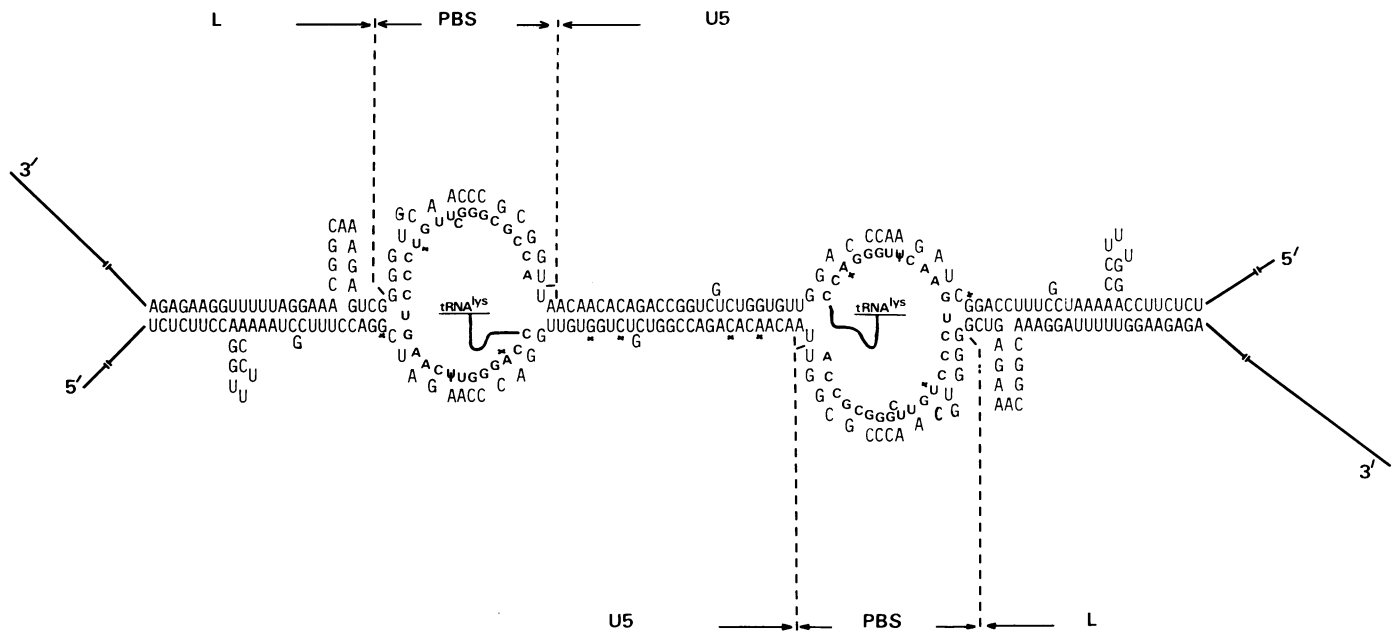


Fig. 6. Possible dimer-linkage structure at the 5' end of E.Tn RNA. This configuration, which is based on the nucleotide sequence, involves two E.Tn RNA molecules and two tRNA molecules. It extends from nucleotide 253 at the 3' end of U5 to nucleotide 372 in the L region and contains the PBS region. x indicates mismatched nucleotides.

sequences are usually at similar positions in retroviral LTRs (Varmus, 1982). An inverted repeat, ending in TGT . . . ACA, borders the E.Tn LTR. Inverted repeats which border retroviral LTRs are involved in the integration of the provirus (Panganiban and Temin, 1983). Duplication of cellular DNA at the junction site has been recognized as a necessary feature of the integration process of transposable elements and retroviruses. A 6-bp duplication brackets the genomic E.Tn sequence in pMAC-2. Integration of retroviruses is accompanied by the loss of two nucleotides from the predicted ends of the linear DNA (Shoemaker *et al.*, 1980; Varmus, 1982). In our case establishing the loss of a dinucleotide from the ends of the LTR upon integration awaits the identification of an unintegrated precursor, if any.

Two bp 3' from the 5'-LTR of E.Tn is a sequence homologous to tRNA<sub>3</sub><sup>Lys</sup>. Retroviruses are known to have a tRNA-binding site at this position which serves as a primer for the synthesis of the minus DNA strand (Weiss *et al.*, 1982; Temin, 1981; Varmus, 1982). Synthesis of a proviral plus strand begins 5' to the 3'-LTR. Although the primer for this event has not yet been identified, the priming site is known to be a purine-rich domain (Van Beveren *et al.*, 1980, 1982; Dhar *et al.*, 1980; Temin, 1981). In fact, a 13-nucleotide purine sequence is located 5' to the E.Tn 3'-LTR. Finally, like many retroviral RNAs, two E.Tn RNAs can potentially form a dimer-linkage structure with two tRNA molecules (Figure 6). A role for such a structure in the encapsidation and possibly the replication of retrovirus RNA has been postulated (Weiss *et al.*, 1982; Watanabe and Temin, 1982; Mann *et al.*, 1983). Thus E.Tn genomic sequences have many of the features necessary to be transcribed and replicated. We have not yet, however, investigated the possible replication of E.Tn RNA in EC cells nor its possible encapsidation into viral-like particles.

RNA-DNA *in situ* hybridization has shown that E.Tn transcription peaks in the pluripotent cell lineages in 3.5–7.5-day embryos and also in extra-embryonic ectoderm layers. Their transcription drops down slowly to a few percent of its peak value

in tissues of older embryos (Brûlet *et al.*, 1985). By Southern and dot blots analysis we have estimated 200–300 E.Tn copies per haploid genome of *Mus musculus*. The cDNA nucleotide sequences presented here show that several of these E.Tn sequences are transcribed in the pluripotent undifferentiated cells. However, even under high stringency conditions of S1 digestion, discrete fragments are generated from regions of the LTR expected to be protected from S1 nuclease by the hybridized RNA. This is strongly suggestive of a homologous population of RNA transcripts even if they originate from different E.Tn genomic sequences. Analysis of the *cis*- and *trans*-acting elements, which restrict the E.Tn transcription to the early embryonic lineages, will add to the understanding of the molecular events in early mouse development.

## Materials and methods

### Cells

The cell lines used in this study were described previously and were cultured under standard conditions (Brûlet *et al.*, 1983).

### Enzymes

Restriction enzymes were purchased from Boehringer Mannheim, BRL or Biolabs and used according to the manufacturers' specifications. T4 DNA ligase and polymerase were from Boehringer Mannheim. Avian myeloblastos virus reverse transcriptase was obtained from Life Sciences Inc., Miami. [<sup>32</sup>P]dNTP (800 Ci/mmol) was from Amersham.

### λgt11 library construction

Double-stranded cDNA was synthesized from 2 μg of PCC4 Aza poly(A)RNA, treated with S1 nuclease to produce blunt ends, protected by reaction with *Eco*RI methylase and ligated to *Eco*RI linkers (Huynh *et al.*, 1984). The cDNA was digested with *Eco*RI and purified by chromatography on Bio Gel A50M in TEN buffer (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 100 mM NaCl). It was then ligated to *Eco*RI-digested phosphatase-treated λgt11 arms (Young and Davis, 1983a). Phage DNA was packaged by using a commercial packaging kit (Amersham) and plated on Y1088 strain (Young and Davis, 1983b).

### Nucleotide sequencing

From an *Eco*RI, *Hpa*II digest of pMAC-2, the 1.2- and 0.8-kb fragments were electrophoretically purified. These fragments were further digested with *Hae*III, the staggered ends were repaired with the Klenow DNA polymerase and *Pst*I linkers were ligated onto the fragments before being inserted into the M13mp8

*Pst*I site. Alternatively, the 1.2- and 0.8-kb fragments were digested with *Sau*3A. Those fragments were ligated to the M13 vector after double digestion with *Acc*I-*Bam*HI, *Acc*I-*Eco*RI and *Bam*HI-*Eco*RI endonucleases. The nucleotide sequencing by the dideoxy-chain termination method was according to BRL's protocol using Biolabs' 15-oligonucleotide sequencing primer.

#### Primer elongation and S1 nuclease mapping

Procedures of RNA extraction, synthesis of single-stranded radioactive DNA probes and S1 nuclease mapping were as previously described (Brûlet *et al.*, 1985). The probe for primer elongation was synthesized from an M13-5'LTR recombinant template with the Klenow DNA polymerase and the 15-oligonucleotide sequencing primer. The synthesized DNA was restricted with an enzyme cutting inside the inserted DNA and, after denaturation, the single-stranded fragment was electrophoretically recovered. The radioactive DNA fragment ( $5 \times 10^4$  c.p.m.) was mixed with 2  $\mu$ g of poly(A) RNA and ethanol precipitated. The nucleic acids were resuspended in 18.4  $\mu$ l of buffer containing 40 mM Pipes pH 6.4, 1 mM EDTA, 80% formamide and heated for 5 min at 75°C. Annealing of the DNA was performed by incubating the sample at 37°C for 5 h in the same buffer containing 0.4 M NaCl. The reaction was stopped with five volumes of 0.3 M sodium acetate (pH 5.2) and ethanol precipitated. The pellet was resuspended into 20  $\mu$ l of a buffer containing 50 mM Tris-HCl pH 8.0, 10 mM dithiothreitol, 8 mM MgCl<sub>2</sub>, 1 mM each dNTP, 60 mM NaCl and 10 units of reverse transcriptase. The mixture was incubated for 1 h at 40°C. Duplicate tubes were digested with S1 nuclease before and after elongation. After ethanol precipitation, the pellets were dissolved in 25  $\mu$ l of 0.3 N NaOH, heated for 30 min at 70°C and the solution was neutralized with 25  $\mu$ l 0.3 N HCl and 0.5 M Tris-HCl pH 8.0. Then 5  $\mu$ g of tRNA were added, the solution extracted once with phenol, once with chloroform and ethanol precipitated.

#### Acknowledgements

We thank Ms D. Boullier for excellent technical assistance. This work was supported by grants from the Centre National de la Recherche Scientifique (UA 1148, ATP 955 153, ATP 955 189), the Fondation pour la Recherche Medicale, the Ligue Nationale Française contre le Cancer and the Fondation André Meyer. One of us (M.K.) was the recipient of a fellowship from the Association pour la Recherche contre le Cancer.

#### References

- Allan, M., Lanyon, W.G. and Paul, J. (1983) *Cell*, **35**, 187-197.
- Brûlet, P., Kaghad, M., Xu, Y.S., Croissant, O. and Jacob, F. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 5641-5645.
- Brûlet, P., Condamine, H. and Jacob, F. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 2054-2058.
- Dhar, R., McClements, W.L., Enquist, L.W. and Van de Woude, G. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 3937-3841.
- Huynh, T.V., Young, R.A. and Davis, R.W. (1984) in Glover, D. (ed.), *DNA Cloning - A Practical Approach*, IRL Press, Oxford, UK, pp 49-78.
- Mann, R., Mulligan, R.C. and Baltimore, D. (1983) *Cell*, **33**, 153-159.
- Messing, J. and Vieira, J. (1982) *Gene*, **19**, 269-276.
- Panganiban, A.T. and Temin, H.M. (1983) *Nature*, **306**, 155-160.
- Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467.
- Seif, I., Khoury, G. and Dhar, R. (1979) *Cell*, **18**, 963-977.
- Shoemaker, C., Goff, S., Gilboa, E., Paskind, M., Mitra, S.W. and Baltimore, D. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 3932-3936.
- Temin, H.M. (1981) *Cell*, **27**, 1-3.
- Van Beveren, C., Goddard, J.G., Berns, A. and Verma, I.M. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 3307-3311.
- Van Beveren, C., Rands, E., Chattopadhyay, S.K., Lowy, D.R. and Verma, I.M. (1982) *J. Virol.*, **41**, 542-556.
- Varmus, H.E. (1982) *Science (Wash.)*, **216**, 812-820.
- Watanabe, S. and Temin, H.M. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 5986-5990.
- Weiss, R., Teich, N., Varmus, H. and Coffin, J. eds. (1982) *RNA Tumor Viruses, (Molecular Biology of Tumor Viruses)*, 2nd ed., published by Cold Spring Harbor Laboratory Press, NY.
- Young, R.A. and Davis, R.W. (1983a) *Science (Wash.)*, **222**, 778-782.
- Young, R.A. and Davis, R.W. (1983b) *Proc. Natl. Acad. Sci. USA*, **80**, 1194-1198.

Received on 25 July 1985