Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 3768–3771, June 1987
Genetics

Nucleotide sequence and evolution of *ETn* elements

(concerted evolution/molecular drive/mouse embryo/endogenous retrovirus)

PIERRE SONIGO*, SIMON WAIN-HOBSON*, LYDIE BOUGUELERET†, PIERRE TIOLLAIS*, FRANÇOIS JACOB‡, AND PHILIPPE BRÛLET‡

*Unité de Recombinaison et Expression Génétique, Centre National de la Recherche Scientifique Unité Associée 271, Institut National de la Santé et de la Recherche Médicale Unité 163; †Unité d'Informatique Scientifique, and ‡Unité de Génétique Cellulaire de l'Institut Pasteur et du Collège de France, Centre National de la Recherche Scientifique Unité Associée 1148, Institut Pasteur, 75724 Paris Cedex 15, France

Contributed by François Jacob, February 9, 1987

ABSTRACT **The *ETn* (for "early transposon") family of long repeated sequences is abundantly transcribed in early mouse embryos from retroviral-like long terminal repeats. Nucleotide sequencing of two elements does not reveal any long open reading frame nor significant homology to retroviral proteins. The genetic polymorphism, monitored by Southern blotting within and across mouse species, reflects a concerted mode of evolution for the *ETn* sequences.**

We previously have identified a family of moderately repeated sequences designated *ETn* for "early transposon" characterized by a specific spatial and temporal pattern of transcription during early mouse embryogenesis. Transcription peaks between 3.5 and 7.5 days, essentially in undifferentiated cells of the blastocyst inner cell mass and embryonic ectoderm, precursor of the germ line (1, 2). An *ETn* element is 5.6-kilobases (kb) long, colinear with ETn RNA, and is delimited by two direct long terminal repeats (LTRs). The 5′ and 3′ ends of the transcribed RNA are located within the LTRs, whose structure is essentially similar to that of retroviral LTRs (3). The developmentally regulated transcription of long repeated sequences has been detected in widely different organisms, including *Drosophila*, sea urchin, and *Dictyostelium* (4). As an approach to studying the possible physiological role of *ETn* transcription in mouse embryos, we have analyzed the genetic structure and polymorphism of the *ETn* family by Southern blotting and nucleotide sequencing. The observed variability within and across mouse species reflects the concerted evolution of the *ETn* elements. Surprisingly, nucleotide sequencing of two elements could not reveal any long open reading frame or significant homology to retroviral proteins.

MATERIALS AND METHODS

Mice. Pure-line mice came from inbred stocks kept at the Institut Pasteur and were a gift from J. L. Guénet. *Mus caroli, Mus cooki, Mus cervicolor, Mus* (or *Pyromys*) *pahari*, and *Mus* (or *Coelomis*) *plathytrix* were gifts from F. Bonhomme.

Southern Blots. DNAs were extracted from liver and spleen and analyzed by standard methods (1). DNAs (15 μg) were digested with *Sau*3A and fractionated on a 1.2% agarose gel. Blots were probed with the nick-translated *ETn* sequence in pMAC-2. Hybridization (50% formamide at 42°C) and washes [0.2× NaCl/Cit (1× = 0.15 M NaCl/0.015 M sodium citrate, pH 7) at 68°C] were under high-stringency conditions.

Nucleotide Sequencing. DNAs (plasmid pMAC2, subclone of phage MG1 or phage MG6) were sonicated, fractionated [600- to 1000-base-pair (bp) fragments], and subcloned into the *Sma* I site of M13mp8 replicative form DNA. Recombinants were identified by *in situ* hybridization using a nick-translated 4.7-kb *Hpa* I fragment of MG1 as probe; 160 and 120 M13 subclones for each *ETn* clone were sequenced by the dideoxynucleotide-termination method as described (5).

RESULTS

Southern Analysis. Southern blots of mouse DNA cut with restriction enzymes that cut at most once in *ETn* sequences and probed with one cloned *ETn* element [pMAC-2 isolated from a BALB/c mouse (1)] usually show smears. This represents a large number of fragments of variable length containing randomly integrated and dispersed *ETn* elements (data not shown). However, when cut with *Sau*3A, only a few bands are detected, which indicates conserved internal restriction sites.

The existence and intensity of these discrete bands in one individual genomic DNA reflect the homogeneity and the amplification of the family within one genome (Fig. 1 *A* and *B*). The intensity of the bands depends on the copy number underlying each band and also on the extent of nucleotide homology with the pMAC-2 probe. We estimated, from the intensity of the bands, that about 200 elements are present per genome in the BALB/c strain. On the other hand, comparison of patterns obtained with various DNAs is indicative of the polymorphism of the family between individuals and species. In contrast to the observed homogeneity within one individual, a clear divergence is observed between mice from different species (Fig. 1 *A* and *B*). These variations in band positions and intensities are reflecting important modifications of the *Sau*3A restriction map across species, associated with either a possible decrease of nucleotide sequence homology with the pMAC-2 probe or a variation in the *ETn* family copy number.

The divergence of the observed patterns fits well within the phylogenetic framework previously established by genetic analysis for the genus *Mus* (6) (Fig. 1*C*). For instance, *M. cooki* and *M. cervicolor*, which share a common ancestor, clearly show a more similar *ETn* restriction map than they do with *M. caroli* (Fig. 1*B*). As well, all the European species and subspecies studied here (*Mus m. domesticus, Mus m. musculus, Mus spretus*, and *Mus spicilegus*) have clearly more similar patterns between themselves than with the rest of the genus (Fig. 1*A*). This good correlation with the phylogenetic tree precludes the hypothesis of the evolution of *ETn* in recent times [a few million years (Myr) at most for the history of *Mus*] by horizontal transfer in a retrovirus-like manner. Moreover, *Mus* (or *Pyromys*) *pahari* and *Mus* (or *Coelomis*) *plathytrix*, who may have shared *stricto sensu* a common ancestor with the genus *Mus* as much as 10 Myr ago, do not contain any *ETn* sequences detectable with the

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.

Abbreviation: LTR, long terminal repeat.

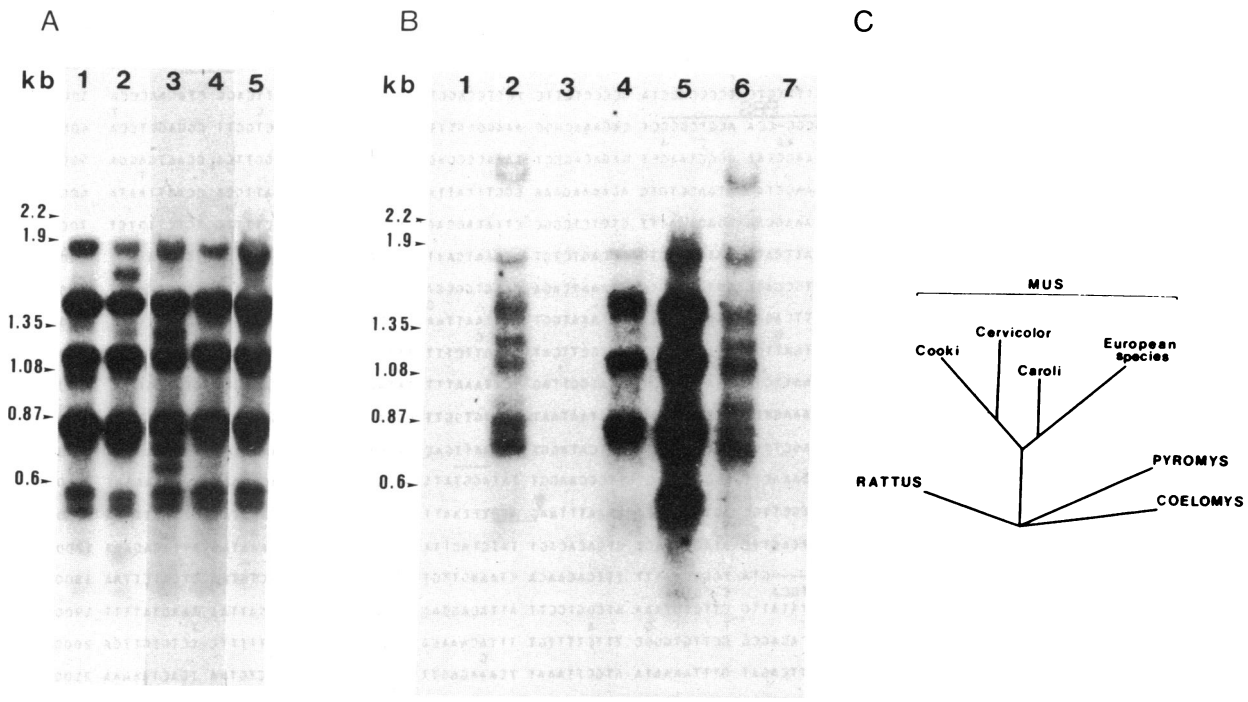

FIG. 1. Phylogenetic distribution and variability of the *ETn* family. (*A* and *B*) Southern analysis of genomic DNAs from various species. Autoradiography was at −80°C for 4 hr (*A*) and for 16 hr (*B*). Calibration of intensity with a reference phage DNA indicates roughly 200 copies in an intense band. (*A*) DNA of European species in lanes: 1, *Mus spretus*; 2, *Mus m. musculus* (PWK inbred strain); 3, *Mus spicilegus*; 4, *Mus m. domesticus* (BALB/c inbred strain); 5, *Mus m. domesticus* (SWR inbred strain). (*B*) DNA of species in lanes: 1, *Pyromys paharii*; 2, *Mus cooki*; 3, *Coelomys plathytrix*; 4, *Mus caroli*; 5, *Mus m. domesticus* (BALB/c strain); 6, *Mus cervicolor*; 7, *Rattus norvegicus*. (*C*) Consensus phylogeny, redrawn from that of Bonhomme (6) for four murid genera analyzed in this work.

pMAC-2 probe under high-stringency conditions (Fig. 1*B*). This result is consistent with Bonhomme's proposal, based on analysis of variations at 28 genetic loci, that *Coelomys* and *Pyromys* are independent genera (6) (Fig. 1*C*).

In more distant genomes (rat, hamster, monkey, and human), *ETn* sequences are not detected under high-stringency conditions. However, a signal is observed in the rat genome under low-stringency conditions (hybridization in 20% formamide/5× NaCl/Cit at 42°C; washing in 2× NaCl/Cit at 45°C; data not shown), but a clear identification of hybridizing DNA will require further analysis.

Nucleotide Sequencing. To determine the genetic structure of *ETn* elements and to appreciate their variability at the nucleotide level, we sequenced two independent clones of *ETn* isolated from a BALB/c mouse DNA genomic library. Clone MG1 is 5544 bp long, is very A+T-rich (62%), and features a low C-G dinucleotide frequency as is typical of eukaryotic DNA. Clone MG6 sequence is 95% complete, lacking only 91 bp from its 5′ LTR and 360 bp from the 3′ LTR. The overall nucleic acid sequence homology between the two studied elements is 94.2% (Fig. 2).

Surprisingly, in both clones, on both DNA strands, no long open reading frame is present (Fig. 3). We have screened these sequences against data banks and particularly all published retroviral or retroviral-like sequences and mammalian repetitive elements and have found no significant similarity. The absence of homology with known retroviral genomic organization or retroviral proteins is in complete contrast with the perfect retroviral features of the *ETn* LTRs. The *ETn* LTR is flanked by a primer binding site complementary to $tRNA_{1,2}^{Lys}$ (8), typically used to prime reverse transcription, and by a polypurine tract, the priming site for retroviral DNA (+)-strand synthesis. The LTRs are bordered by inverted repeats containing the conserved T-G . . . C-A dinucleotides. Direct repeats bracket the entire element as a usually observed consequence of retroviral integration (9).

We found that the best alignment of the LTR sequence was with that of the D-type retrovirus Mason–Pfizer monkey virus, which is 67% homologous, with the same binding site, cap, and polyadenylylation sites and a very similar polypurine tract (5). All these data strongly suggest that *ETn* LTRs are indeed of retroviral type.

DISCUSSION

The existence of genetic elements such as *ETn* raises many questions as to their origin, their possible function or influence on evolutionary or developmental processes, and their mode of evolution. The *ETn* could have been derived from a retrovirus involving the degeneration of the internal sequences or the recombination between two retroviruses or solitary LTRs resulting in the "capture" of genomic DNA.

Diverse retroviral-related sequences have already been characterized in the mouse genome. Listed by increasing order of genetic content and independence from the host genome, they include solitary LTRs, virus-like VL30 sequences, intracisternal A-type particle (IAP) sequences, and finally retroviral genomes (see ref. 9 for a review). *ETn* sequence could be classified between solitary LTRs and VL30. In Temin's hypothesis on the origin of retroviruses from cellular moveable genetic elements (10), retroviruses are generated from an ancestral gene by successive cycles of transcription, reverse transcription, and integration. By adding genetic information, the cycles finally gave rise to a proviral-like element. The existence of *ETn* elements with their typical retroviral features and unstructured internal domains might give credence to such an ongoing phenomenon. In this context, the *ETn*, which could be both a defective descendant and a potential progenitor of a retrovirus, would reflect the bidirectional aspect of the process.

ETn sequences were isolated during the course of studies on the early mouse embryogenesis. The absence of a long

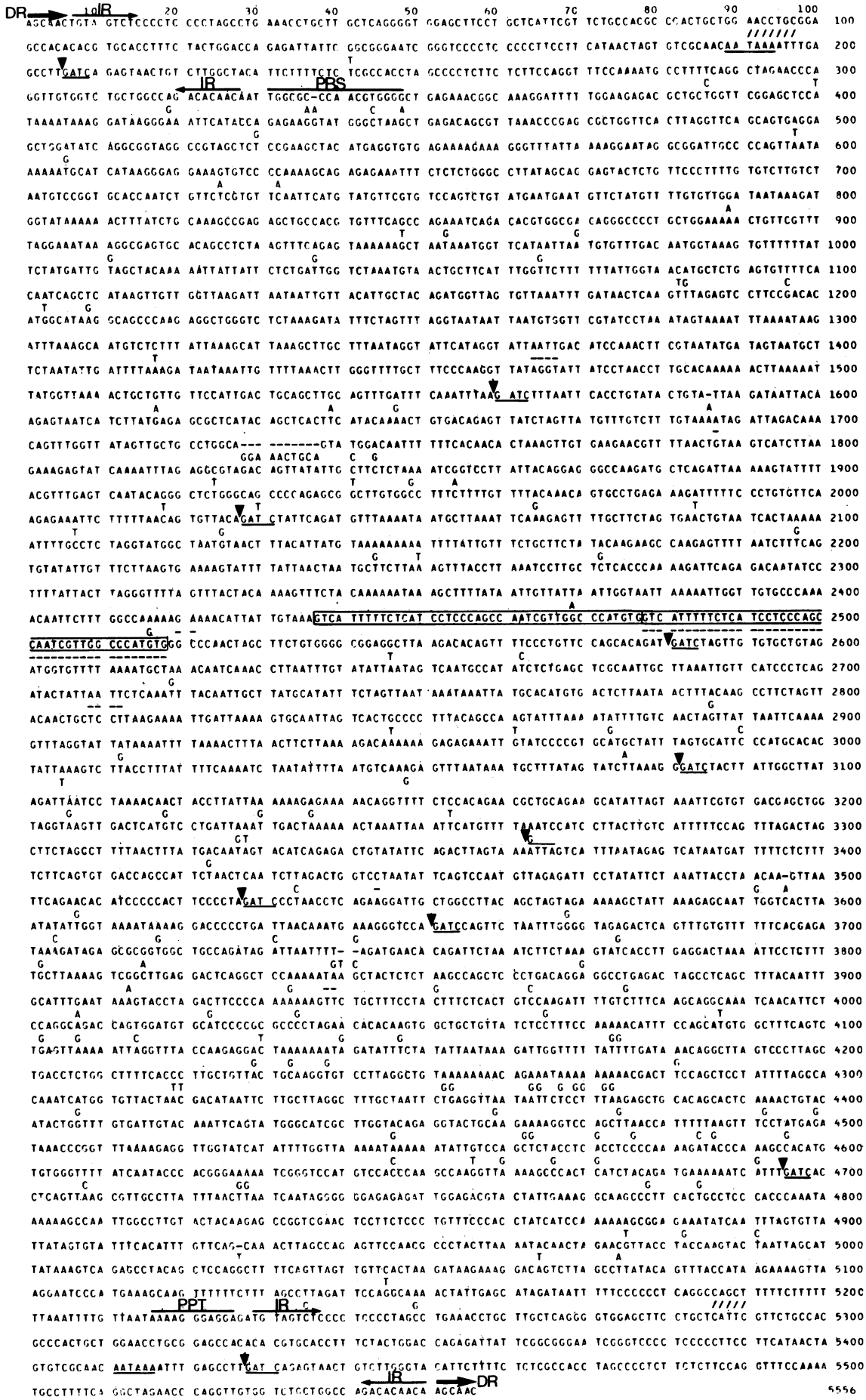

FIG. 2. *(Legend appears at the bottom of the opposite page.)*

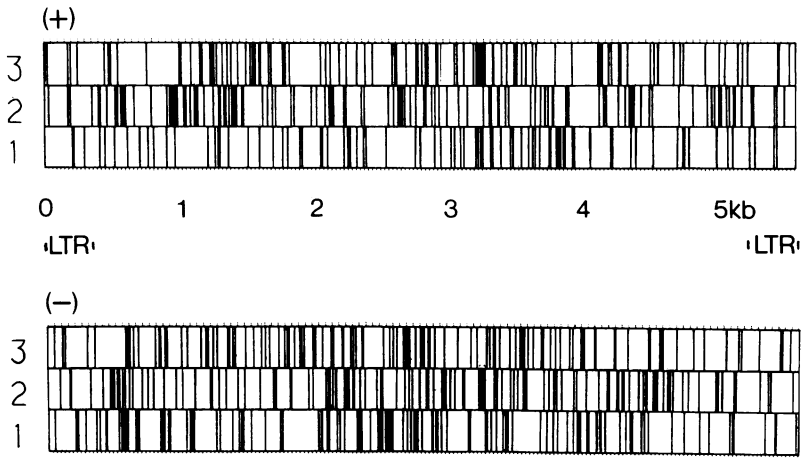

FIG. 3. Coding potential of the *ETn* element. The sequence of MG1 was read in all three reading frames (1, 2, and 3) scoring only stop codons, represented by vertical bars. Both strands of the element are shown (+) and (−). No significant nucleic or amino acid sequence homology was found between any of the small open reading frames and those in data banks, especially retroviruses, retrovirus-like, or repetitive elements. Similar results were found for clone MG6.

open reading frame suggests that *ETn* could be only "selfish" or "parasitic" DNA (11, 12). However, we cannot rule out the possibility that they encode small proteins that may have important regulatory effects [like, for example, the transactivating genes of human immunodeficiency virus (13)]. Furthermore, the evolutionary dynamics of the family may play a role in the genome during embryogenesis or evolution of the species.

Remarkably enough, the pattern of variation of the *ETn* family parallels the phylogeny of mice and fits the definition of concerted evolution—that is, the homology between elements of the same family within a species is higher than that between members of different species. The primary mechanism underlying this concerted mode of evolution may involve the selective amplification of one or a few copies in each species. However, an additional process of "homogeneization" might be required for eliminating old elements and fixing the amplified variants in the population. The necessity of such a process led Dover to introduce the notion of molecular drive (14), which is based on a variety of genomic asymmetric turnover mechanisms, independent of natural selection. In the case of the interspersed *ETn* family, which is abundantly transcribed in early embryonic cells, and because of the presence of a typical retroviral LTR, we think that transcription, reverse transcription, and integration, possibly by homologous recombination at already occupied sites, might be involved in the observed amplification and homogeneization. This postulated process would include a simple "molecular selection" since essentially one selected element of the family has spread in each species. Small differences in the efficiency of LTR sequences for promoting transcription, reverse transcription, and/or integration could be a basis for this selection.

Clearly, detailed information has to be gained on the molecular mechanisms underlying concerted evolution, molecular drive, and their eventual relationships with phylogeny and ontogeny. Because of its relatively low copy number and high level of transcription in undifferentiated cells, the system provided by the *ETn* family in the genus *Mus* is amenable to experiments.

We thank D. Boullier for technical assistance, Drs. F. Bonhomme and J. L. Guenet for the gift of mice and their critical comments on the manuscript, and Drs. M. Emerman and J. Weissenbach for useful advice. This work was supported by grants from the Association pour la Recherche sur le Cancer, the Centre National de la Recherche Scientifique (ATP 955189), the Fondation pour la Recherche Médicale, and the Institut National de la Santé et de la Recherche Médicale (851004).

1. Brûlet, P., Kaghad, M., Xu, Y.-S., Croissant, O. & Jacob, F. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 5641–5645.
2. Brûlet, P., Condamine, H. & Jacob, F. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 2054–2058.
3. Kaghad, M., Maillet, L. & Brûlet, P. (1985) *EMBO J.* **4,** 2911–2915.
4. Davidson, E. H. & Posakony, J. W. (1982) *Nature (London)* **297,** 633–635.
5. Sonigo, P., Barker, C., Hunter, E. & Wain-Hobson, S. (1986) *Cell* **45,** 375–385.
6. Bonhomme, F. (1986) *Curr. Top. Microbiol. Immunol.* **127,** 19–34.
7. Alizon, M., Wain-Hobson, S., Montagnier, L. & Sonigo, P. (1986) *Cell* **46,** 63–74.
8. Raba, M., Limburg, K., Burghagen, M., Katze, J. R., Simsek, M., Heckman, J. E., Rajbhandary, U. L. & Gross, H. J. (1979) *Eur. J. Biochem.* **97,** 305–318.
9. Temin, H. M. (1985) *Mol. Biol. Evol.* **2,** 455–468.
10. Temin, H. M. (1980) *Cell* **21,** 599–600.
11. Doolittle, W. F. & Sapienza, C. (1980) *Nature (London)* **284,** 601–603.
12. Orgel, L. E. & Crick, F. H. C. (1980) *Nature (London)* **284,** 604–607.
13. Chen, I. S. V. (1986) *Cell* **47,** 1–2
14. Dover, G. (1982) *Nature (London)* **299,** 111–117.

FIG. 2 *(on opposite page).* DNA sequence of two *ETn* elements. The upper line gives the sequence for clone MG1, while the second line gives that of clone MG6 in which only the differences with respect to the MG1 sequence are shown. The slashes (//) denote the ends of our MG6 sequence, which is 95% complete. The sequence is annotated according to features of a typical retroviral LTR:DR, direct repeat (of cellular origin); IR, inverted repeat; PBS, tRNA primer binding site; and PPT, polypurine tract. *Sau*3a restriction sites are underlined and marked by triangles. A long direct repeat of 41 bp in the MG1 sequence with respect to the MG6 sequence is boxed. Such direct repeats are usually observed as a source of variability, especially in retroviruses (see ref. 7 for an example). Each base was sequenced on average 6.3 (MG6) or 5.7 (MG1) times. The base composition of the MG6 (+)-strand is: T, 31.7%; C, 20%; A, 30.3%; and G, 18%, with A+T = 62%. The two sequences are 94.2% identical, and the number of transitions and transversions are 137 (2.7%) and 88 (1.7%), respectively.