Volume 16 Number 2 1988 Nucleic Acids Research

The F-type 5' motif of mouse L1 elements: a major class of L1 termini similar to the A-type in organization but unrelated in sequence

Richard W.Padgett*, Clyde A.Hutchison III and Marshall H.Edgell

Curriculum in Genetics, Department of Microbiology and Immunology, Program in Molecular Biology and Biotechnology, University of North Carolina, Chapel Hill, NC 27514, USA

Received September 4, 1987; Revised and Accepted December 14, 1987

ABSTRACT

It has previously been shown that the L1 family in the mouse (L1Md) contains two alternative 5' ends called the A- and F-type sequences (1,2). We show here that the F-type element is a major class of murine L1 elements and report on the details of organization of the 5' motif of these F-type elements. Although the A- and F-type 5' sequences share no detectable sequence homology the organization of an F-type 5' end is strikingly similar to that of an A-type. That is, the F-type 5' sequences consist of a tandem array of a small number of 206 bp monomers while the A-type 5' motif consists of a tandem array of 208 bp monomers. All of the A-type elements characterized to date have a truncated monomer at the 5' end of the array. Many of the F-type elements are also terminated at the 5' end by a truncated copy but unlike the A-type elements some F-type elements terminate with a monomer which is within a few nucleotides of being complete. In addition the F-type consensus sequence, in contrast to the A-type sequence, shows homology (70%) to the body of the L1Md starting at the position where the monomer joins the rest of the L1 element.

INTRODUCTION

L1 is one of the most abundant dispersed repetitive families in mammals (3,4). The longest examples in the mouse and human are 6-7 kb in length. Most members are bounded by short direct repeats and contain a poly-A rich tract at the 3' end which is usually interpreted to mean that an RNA intermediate is involved in their dispersal throughout the genome (5,2). Most individual members in the genome carry polar truncations of varying length at their 5' ends (6). About 10,000 of the 100,000 L1 elements appear to be full length (M. Comer, manuscript in preparation).

L1 transcripts from rodents and primates have been identified (7,8,9) and shown to be products of RNA polymerase II (10). L1 elements in both rodents and primates have been shown to contain open reading frames (ORFs) (11,12). Martin et al. (12) showed that one of the open reading frames in the L1 family was evolving under selection for protein function by examining the replacement and synonymous site substitutions in a comparison of primate and rodent sequences.

© IRL Press Limited, Oxford, England.

The sequences of a full length murine L1 element, L1Md-A2, has been reported (1). 13 bp direct repeats define the limits of the L1Md-A2 element. This mouse sequence contains two long open reading frames, one 1137 bp and the other 3900 bp in length. Four and two-thirds copies of a 208 bp repeating unit, which we have called the A type monomer, are found at the 5' end of the element. Examination of the 5' end of a second long copy of L1, L1Md-A13 (13), revealed the presence of one and two-thirds copies of the A-type monomer suggesting that variable numbers of these A repeats are associated with the 5' end of L1 and that the 5'-most copy is usually truncated leaving a two-thirds copy (1). This repetitive A-type monomer is dispersed throughout the genome and is associated "exclusively" with the L1 family (M. Comer, manuscript in preparation).

It has been proposed that the A-type monomer contains a promoter and that transcription initiates about two-thirds of the way upstream from the 3' end of the monomer (1). Each RNA molecule that undergoes retroposition into the genome would then possess a two-thirds copy of the A-type monomer at its 5' end.

Fanning (2) reported the sequence of several L1/non-L1 5' junction fragments from mice. Alignment of the sequences from two of those clones indicated the presence of repetitive sequence not previously reported to be part of the L1 structure. He concluded that the point where those sequences lost homology was the 5' end of mouse L1 elements. Comparison of this sequence to that of the L1 element sequenced by Loeb et al., indicates that the 5' sequence reported by Fanning (2) was attached to the body of L1 at the same position that the A-type monomers are attached but shares no homology with the A-type monomer. This suggests that these different 5' sequences represented alternative 5' ends of mouse L1 elements. We call this alternative 5' sequence in L1 elements the F-type sequence. This unusual finding, that the repetitive L1 family has two alternative 5' ends, has led us to examine the structure and organization of the F-type 5' sequence in more detail. Using sonicated total genomic DNA from the mouse, we constructed an M13 library which was subsequently probed to isolate L1 clones with F monomers. This method of preparing clones avoids the prerequisite of a detailed restriction map to generate clones and does not bias the library against polymorphic subfamilies of the L1 repeats. Here we report the organization and structure of the 5' ends of the F-type L1 elements as deduced from these clones.

MATERIALS AND METHODS

Preparation of Genomic DNA for Cloning

High Molecular weight DNA from the laboratory mouse, Mus domesticus, was prepared essentially as described by Blin and Stafford (14). After phenol and chloroform extraction the DNA was collected from a CsCl gradient according to Maniatis et al. (15). The preparation of DNA for cloning was essentially as described in Deininger (16) but with the modifications of Bankier and Barrell (17). 5µg of total genomic DNA were sonicated in a microfuge tube clamped 1 mm above the probe of a Heat Systems Ultrasonics Cup Horn Sonicator. Four bursts of 40 seconds each on maximum power rendered the DNA an average of 600 bp in length. The water in the cup horn sonicator was changed after every other burst to prevent overheating the DNA. This mixture was end repaired using 20 units of the large fragment (Klenow) from E. coli DNA polymerase I. After electrophoresis on a 1.5% (w/v) agarose gel, DNA fragments 300-800 bp in length were eluted and recovered from DE81 DEAE-cellulose paper as described by Dretzen et al. (18). 50-100 ng of these DNA fragments were blunt-end ligated into 20 ng of M13 mp18 which had been cleaved with Sma I and alkaline phosphatased in a total reaction volume of 10µl (17).

Plating and Screening the Random M13 Clones

Ligations of end-repaired mouse DNA and Sma I cleaved/alkaline phosphatased M13 were transfected into competent JM107 cells as described by Hanahan (19). Conditions were established so that about 500 plaques were plated on each 100mm petri plate. About fifty percent of the phage contained inserts as indicated by lack of beta galactosidase activity when plated with X-gal and as determined by sequencing randomly selected phage. To select clones containing the 5' end of F-type elements, a 201 bp Eco RI/Bam HI fragment unique to the 5' end of the F-type L1 elements was isolated from the plasmid clone (pBfL-5; described in (2) and used as a probe. This fragment has been cloned into M13 mp10 by M. Comer (our laboratory) and is called TF-10. Nick translation, plaque lifts and hybridization were performed as described by Jahn et al. (20). Under these conditions about 1 phage with an insert in 1000 was positive for the F-type sequence.

Sequencing Protocols

All sequences were determined by the dideoxy method of Sanger and Coulson (21) as modified by Bankier and Barrell (17). Reactions using ^{35}S

labeled nucleoside triphosphates were done in microtitre plates using a Hamilton stepping device to dispense samples onto the sides of the solutions in each well. Samples were heated at 80°C for 15 minutes to denature and concentrate them. Samples were electrophoresed on 20 x 50 cm buffer gradient gels (22) or on nongradient 20 x 60 cm gels for extended sequence determination of clones. Up to 12 samples (48 lanes) were loaded on each gel and 150-250 bases determined from each.

Computer Analysis of Sequence Data

A series of programs to analyze DNA sequences have been developed in our laboratory to work with a Z-80 based microcomputer operating under CP/M. Among them are a sequence editor (SED), a sequence display (FIGMAKER), a restriction site search program (ALLSITES)(23), a dot matrix display program and a sequence alignment program (DIAGSRCH, DIAGPLOT and SALT)(24), SEQINP, a simple and convenient method for reading and storing sequence gel information developed by C.A. Hutchison (25), allows one to input data by typing directly from sequence autoradiographs into a portable computer (TRS-80, Model 100). The program assigns each base a specific audio tone to aid accurate typing of the sequence. The sequence files can be concatenated into one long file using BATCAT for transferral to another computer for data analysis.

RESULTS

BALB/c genomic DNA fragments were prepared for the mini-library construction by shearing the DNA via sonication and repairing the ends for blunt end ligation with the Klenow fragment of *E. coli* DNA polymerase I. This method has the obvious advantage that by not using restriction cleavage of the input material it does not bias the yield of clones for family members within a particular range of divergence. Fragments 300-800 bp were ligated into M13 mp18 to produce a small library for screening. The *Eco* RI/*Bam* HI insert fragment from TF-10 (methods) containing the F-type sequence was nick translated and hybridized to plaque lifts of the M13 library of random genomic fragments. About half of the plaques on the plate contained phage without inserts as judged by beta-galactosidase activity. Approximately 1 clone per 1000 clear plaques hybridized to the F-type sequence. Hybridizing clones were picked and replated for purification in order to prepare DNA for sequencing.

F-Type Sequences Exist in Tandem Arrays.

The sequences generated from the clones carrying the F-type sequences

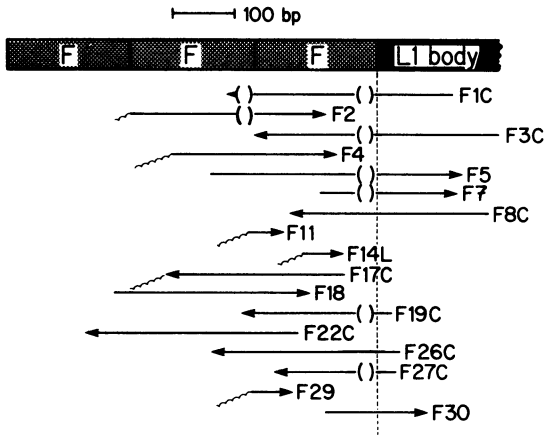

Figure 1. Position of random F-type clones. Each of the random F-type clones was positioned relative to a consensus array of three copies of the F-type repeat. The sequence of the body of L1Md-A2 is attached to the F-type arrays as shown. The direction of the arrows indicate the orientation of the clones. The name of each clone is shown beside the arrows. The length of each arrow represents the length of the sequence determined from each clone. The wavy lines represent non-homology to the F-type consensus sequence. Parentheses indicate a 27 base pair deletion relative to the consensus sequence.

were all aligned to each other (Figure 1 and 2) using the sequence alignment tool SALT. The majority of these clones contained an internal repetitiveness indicating that the F-type 5' sequence occurs as a tandem array similar to what was found with the 5' end of A-type L1 elements. Sequence from F-type clones containing part of the body of L1Md were aligned with sequence from an A-type L1 element, L1Md-A2 (1). The position at which L1Md-A2 and the F-type clones lost homology was used to define the 3' end of the F-type monomer. The 5' end of the F-type monomer was determined by noting the base at which the internal repetitiveness within the F-type clone began. The size of the consensus F-type monomer determined in this manner, is 206 bp which is surprisingly similar to the length of the A-type monomer which is 208 bp.

The F-type monomer contains a 27 bp element which is duplicated in most of clones (positions 150 to 203, Fig. 2). In several clones, there is a deletion of one of the 27 bp elements. The dominant process appears to be that of deletion relative to an ancestral version with two 27 bp sequences since the two sequences carry a distinguishing polymorphism at position 9 within the 27mer (residues 158 and 185, Fig. 2) which is preserved in all

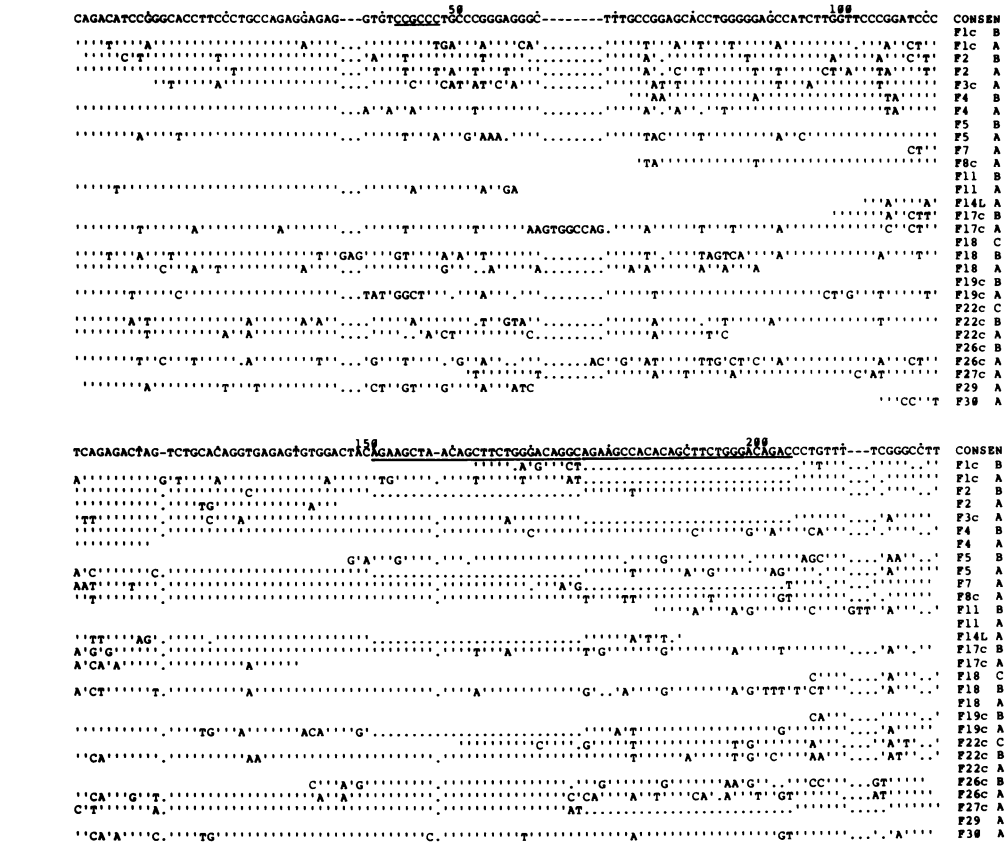

Figure 2. Alignment of the F-type monomers. Sequenced clones from the random library were aligned according to homology. Clone names appear on the right side of the figure. The capital letter (A, B, or C) after the name indicates the proximity of the subunit to the body of the L1 element, "A" being the closest to the body, "C" being the farthest. Pads, designated as "." were introduced to maximize the alignment. The consensus sequence appears at the top of the figure. Gaps ("-") are present on the consensus to accommodate insertions within individual monomers with the exception of the gap at position 158. The gap at this position is to put the two 27 bp repeats (underlined and running from 150 to 176 and 176 to 203) into maximum homology. Bases within individual monomers which differ from the consensus are indicated by the appropriate base (A, G, C, or T). Bases which are identical to that of the consensus are indicated by the "'" marks. The promoter core sequence (32) is underlined and is from bases 42-47.

members with two copies of this sequence. Six of nine of the F-type monomers just 5' to the body of the L1 repeat contain one copy of the 27 bp dimeric structure (Fig. 1). All the copies of the F-type monomer that are

not immediately adjacent to the body of L1Md contain two copies of this dimeric sequence except F1C and F2. This dimeric sequence is not found in the A-type monomer although there is an eight bp region in the middle of the A-type monomers which is dimeric in some examples and monomeric in others (1). The position, length and sequence of this 8 bp feature is different from the dimeric sequence found in the F monomer.

The consensus sequence (Figure 2) derived from these sequences contains six CG dinucleotides. The sites for these dinucleotides show considerably more variation than most other sites in the F monomer which is presumably due to the high rates of methylation of the C within the target CG dinucleotide (26 and Royal McGraw, manuscript in preparation). At saturation one expects all of the sites to be filled with either TG or CA.

Structure of the 5'-Most monomer in the F-type tandem arrays.

Six of the seventeen F-type clones carry sequence at their 5' ends which show no homology to the F-type monomer and presumably represents the non-L1 sequences into which the L1 elements were inserted. These six non-L1 sequences show no significant homology to each other. The 5' end-points of the F-type elements defined by these points of non-homology cluster into two areas. Three of the clones, F4, F14L and F17C terminate at a point about two-thirds of the way from the 3' end of a monomer unit (Figures 1 and 2). The A-type L1 elements terminate at a very similar location, that is, two-thirds of the way from the 3' end of the A-type monomer. About 40-50 bases 5' to where these three clones lose their homology to the consensus F-type sequence is a sequence (positions 42 to 47, Fig. 2) homologous to the proposed promoter in several constitutively expressed genes such as dihydrofolate reductase (27,28) HMG-CoA (28), adenosine deaminase (29) and hypoxanthine phosphoribosyl transferase (30). Transcripts appear to be initiated 30-40 bp 3' of the core CCGCCC in at least one of these characterized systems (28). We do not find any matches within the F-type monomer sequence to the alternative pol II promoter sequence -CCAAT -30 bp -ATA-.

The other three clones which carry non-L1 sequence at their 5' ends, F2, F11 and F29, appear to terminate within a few bases of each other (Figure 1) at a point very close to the 5' boundary of the F-type monomer. The sequences 5' to this point in the three clones were compared to each other, to F and A monomers, and to the body of the L1 element at various criteria. No homology was detected.

Homology between the F-type monomer and the body of L1Md.

Low stringency computer searches were done among the A, F and L1Md sequences. No detectable homology between the A and F-type monomers was found using various search criteria. DNA hybridization experiments between the F and the A-type monomer clones also showed no homology (data not shown). No detectable homology between the A-type monomer and the remainder of the L1Md consensus sequence was found by computer search. However, we were able to find homology between the F-type monomer and the body of the L1Md sequence. The 5' end of the F-type monomer shows 76% homology to a 40 bp region within the first 70 bp of the body of L1Md.

DISCUSSION

Randomly cloned examples of F-type L1 5' sequences have been isolated from M13 libraries. As such clones were abundant we can conclude that the F-type of L1 sequence must represent a major class of L1 element within the mouse genome. DNA sequences from these clones indicates that the F-type and A-type 5' motifs to L1Md elements share an analogous organization consisting of tandem arrays of a monomer about 200 bp in length. The A-type monomer is 208 bp and the F-type monomer is 206 bp. This conserved length to the monomer is puzzling. It suggests a structural boundary condition acting on the monomer length but not on the total L1 element length since the elements can have variable numbers of the monomers. The two monomers do not share extensive sequence homology. Both tandem arrays attach to the body of the L1 element at exactly the same base. The range or average number of copies of the monomers in the tandem arrays of long L1 repeats is not known for either type of L1 element. The longest array of A-type monomers found so far contains 4 and 2/3 copies (1). The longest array found among the M13 clones of the F-type L1 elements contain parts of three F-type monomers (clones F18 and F22C; Fig. 1), but the method of cloning into M13 which we employed would limit the number of copies which could be contained in each F-type clone. Unequal crossing-over events between different L1 elements may give rise to a varying number of copies of these 5' monomers.

Several sequences from the 5' end of L1 elements in primates have been collected (31). The 5' end of the human L1 element has been established by identifying several clones whose sequences diverge from one another at a common base. Surprisingly, the human L1 appears to have no repeating structure similar to the A- or F-type sequences at its 5' end. Since the L1 elements in the mouse have maintained a common 5' motif with two quite

different sequences it would seem that the 5' motif must be important to the preservation of these sequences within the genome. Hence, it is not clear to us how such a motif could come to be absent in the human version of the L1 element.

The 5' end of the F monomer has homology to the body of L1Md. This homology suggests a model to explain the existence of two different types of 5' ends in the L1 population. An unequal crossing-over event between an F-type monomer or multimer preexisting in the genome but not associated with L1 and an A-type L1 element could occur in the region of shared homology (Fig. 3). If the crossing over occurred at the 5' end of the homologous region of L1 then the body of the L1 element would retain its original sequence and remain 76% homologous to the F repeat while acquiring a new 5' end. This new structure would have to become fixed in the population and spread throughout the genome to become as abundant as it is today. Since most F monomers are currently associated with L1Md, as assayed by lambda library plaque screening (M. Comer, manuscript in preparation), this would argue that the F monomer was relatively rare in the genome prior to its association with L1 or that it has since been lost from the genome as an isolated monomer. Once captured, the F monomer may have given L1 some selective advantage in a particular cell or tissue and then spread through the genome. All of the A-type L1 elements characterized so far are terminated by a two-thirds copy of the 5'-most monomer (1). Loeb et al

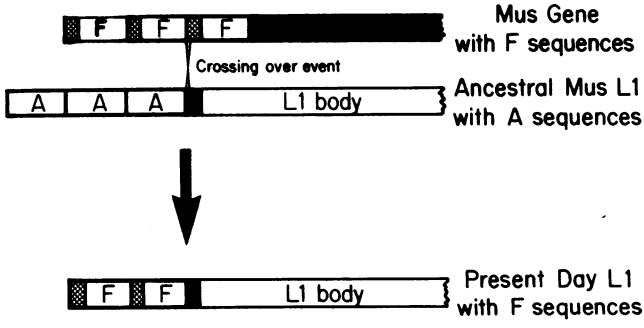

Figure 3. Diagram showing model of origin of two 5' ends. The upper portion of the figure shows an unidentified non-repetitive gene that contained a 5' array of the F repeat. The 5' portion of each F repeating unit (hatched area) has homology with the 5' portion of the body of L1Md (black box). A crossover event between these two genes, as indicated, would result in an L1Md repeat containing the 5' end of the unidentified mouse gene containing the F repeats.

proposed that this was due to the A-type monomer containing a house-keeping type of pol II promoter near the truncation point. Transcription initiation at this point would generate a transcript, which upon retroposition would give rise to an Ll element with a truncated 5' monomer. The fact that the structure of the 2/3 truncated F-type elements is so similar to the structure of the A-type elements strongly suggests that a common mechanism underlies the truncation of both types of elements within the tandem arrays. Since the sequences of the A- and F-type monomers show no detectable homology but do share the consensus -CCGCCC- pol II type promoter (32) such a model is attractive. The endpoints of the F-type elements that truncate at the end of a full repeating unit are presumably produced by some other mechanism.

ACKNOWLEDGEMENTS

We thank M. Agostino for manuscript review and Joyce Bradshaw for preparing the manuscript. This research was supported by Public Health Service Grant GM 21313.

*Present address: Department of Cellular and Developmental Biology, Harvard University, Cambridge, MA 02138, USA

REFERENCES

1. Loeb, D.D., Padgett, R.W., Hardies, S.C., Shehee, W.R., Comer, M.B., Edgell, M.H. and Hutchison III, C.A. (1986) Mol. Cell. Biol. **6**, 168-182.
2. Fanning, T.G. (1983) Nucl. Acids Res. **11**, 5073-5091.
3. Bennett, K.L., Hill, R.E., Pietras, D.F., Woodworth-Gutai, M., Kane-Haas, C., Heath, J.K. and Hastie, N.D. (1984) Mol. Cell. Biol. **4**, 1561-1578.
4. Sun, L., Paulson, K.E., Schmid, C.W., Kadyk, L. and Leinwand, L. (1984) Nucl. Acids Res. **12**, 2669-2690.
5. Voliva, C.F., Jahn, C.L., Comer, M.B., Hutchison III, C.A. and Edgell, M.H. (1983) Nucl. Acids Res. **11**, 8847-8859.
6. Voliva, C.F., Martin, S.L., Hutchison III, C.A. and Edgell, M.H. (1984) J. Mol. Biol. **178**, 795-813.
7. Fanning, T.G. (1982) Nucl. Acids Res. **10**, 5003-5013.
8. Kole, L.B., Haynes, S.R. and Jelinek, W.R. (1983) J. Mol. Biol. **165**, 257-286.
9. Lerman, M.I., Thayer, R.E. and Singer, M.F. (1983) Proc. Natl. Acad. Sci. USA **80**,3966-3970.
10. Shafit-Zagardo, B., Brown, F.L., Zavodny, P.J. and Maio, J.J. (1983) Nature **304**, 277-280.
11. Potter, S.S. (1984) Proc. Natl. Acad. Sci. USA **81**, 1012-1016.
12. Martin, S.L., Voliva, C.F., Burton, F.H., Edgell, M.H. and Hutchison III, C.A. (1984) Proc. Natl. Acad. Sci. USA **81**, 2308-2312.
13. Shehee, W.R., Chao, S.-F., Loeb, D.D., Comer, M.B., Hutchison III, C.A. and Edgell, M.H. (1987) J. Mol. Biol. (in press).

14. Blin, N. and Stafford, D.W. (1976) Nucl. Acids Res. **3**, 2303-2308.
15. Maniatis, T., Hardison, R.C., Lacy, E., Lauer,J., O'Connell, C., Quon, D., Sim, G.K. and Efstratiadis, A. (1978) Cell **15**, 687-701.
16. Deininger, P.L. (1983) Anal. Biochem. **129**, 216-223.
17. Bankier, A.T. and Barrel, B.G. (1983) In, Flavel, R.A. (ed), Techniques in Nuclei Acid Biochemistry, Elsevier Scientific, Limerick, Ireland, Vol. B5, pp. 1-34.
18. Dretzen, G., Bellard, P., Sassone-Corsi, P. and Chambon,P. (1981) Anal. Biochem. **112**, 295-298.
19. Hanahan, D. (1983) J. Mol. Biol. **166**, 557-580.
20. Jahn, C.L., Hutchison III, C.A., Phillips, S.J., Weaver, S., Haigwood, N.L., Voliva, C.F. and Edgell, M.H. (1980) Cell **21**, 159-168.
21. Sanger, F. and Coulson, A.R. (1978) FEBS Letters **87**, 107-110.
22. Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) Proc. Natl. Acad. Sci. USA **80**, 3963-3965.
23. Lautenberger, J.A., White, C.T., Haigwood, N.L., Edgell, M.H. and Hutchison III, C.A. (1980) Gene **9**, 213-231.
24. White, C.T., Hardies, S.C., Hutchison III, C.A. and Edgell, M.H. (1984) Nucl. Acids Res. **12**, 751-766.
25. Hutchison III, C.A. (1986) Nucl. Acids Res. **14**, 1917.
26. Bains, W. (1986) J. Mol. Evol. **23**, 189-199.
27. Yang, J.K., Masters, J.N. and Attardi, G. (1984) J. Mol. Biol. **176**, 169-187.
28. Reynolds, G.A., Basu, S.K., Osborne, T.F., Chin, D.J., Gil, G., Brown, M.S., Goldstein, J.L. and Luskey, K.L. (1984) Cell **38**, 275-285.
29. Valerio, D., Duyvesteyn, M.G.C., Dekker, B.M.M., Weeda, G., Berkvens, T.M. van der Voorn, L., van Ormondt, H. and van der Eb, A.J. (1985) EMBO J. **4**. 437-443.
30. Melton, D.W., Konecki, D.S., Brennand, J. and Caskey, C.T. (1984) Proc. Natl. Acad. Sci. **81**, 2147-2151.
31. Grimaldi, G., Skowronski, J. and Singer, M.F. (1984) EMBO J. **3**, 1753-1759.
32. Reynolds, G.A., Basu, S.K., Osborne, T.F., Chin, D.J., Gill, G., Brown, M.S., Golstein, J.L. and Luskey, K.L. (1984) Cell **38**, 275-285.