Conservation in the 5′ region of the long interspersed mouse L1 repeat: implications of comparative sequence analysis

Estelle Mottez, Peter K.Rogan and Laura Manuelidis

Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06510, USA

Received 29 July 1985; Revised 11 December 1985; Accepted 25 February 1986

ABSTRACT

A clone of 7.1kb corresponding to the mouse L1 interspersed repeat family was selected for homology to a human interspersed repeat. This clone fairly represents mouse genomic members. Mapping of the clone revealed one common element at both the 5' and 3' ends in a head to tail arrangement, suggesting that at least some long L1 family members are tandemly arranged; genomic studies confirmed the unexpected tandem arrangement of a minor proportion of L1 members. A short SmaI tandem repeat appears to define the 5' end of most L1 family members. SmaI repeats may maintain, via a recursive regulatory function, the transcriptional viability of L1 members after retroposition events. A 2.5kb portion of the mouse L1 repeat that has not been previously sequenced is presented. It is 55-70% homologous to a corresponding portion of the human KpnI repeat family. Comparative sequence analysis revealed that one common open reading frame may conserve potential coding function across species. A second open reading frame bears an asymmetric distribution of codon replacements unlike both genes and pseudogenes. This latter feature could be consistent with a proposed chromosome organization function that is unrelated to peptide expression.

INTRODUCTION

Eukaryotic genomes contain a large proportion of interspersed repetitive DNA sequences (1) which may be classified according to their length and lack of tandem repetitions. One prominent family of long interspersed repeats occurs in rodents and has recently been renamed the L1 family (2); this repeat had previously been described with respect to shorter abundant restriction fragments [e.g. BstNI (3), MIF-1 (4-6), BamHI (7), Bam5 (8), and "R" repeats (9)]. These repeats are frequently linked together in larger genomic elements of at least 7 kb (2,6) or even 9 kb in length (10). However, sequences from the 3' end of the L1 repeat are more abundant than are those at the 5' end (2) suggesting truncation of at least some genomic family members.

Nuclear transcripts corresponding to more 5' regions of this family, both in mouse (11) and homologous human LINES (12-16) have been detected.

No sequencing studies of the more 5' region in mouse have been reported. In mice, BstNI 1.5kb and 1.7kb restriction fragments were initially described as two non-homologous repetitive sequences that are dispersed on all chromosomes (3). Both of these fragments are part of the L1 family. These sequences appear to be conserved in evolution, since the mouse BstNI 1.5kb repeat hybridized to the human HindIII 1.9kb repeat (17,18), which represents a more 5' region of the KpnI primate "LINE" family (19,20), and the more 3' BstNI 1.7kb repeats (equivalent to the mouse EcoRI 1.3kb or "MIF-1" repeat) also hybridized with human DNA (17). Sequence studies of cloned examples from mouse L1 and primate Kpn LINES have also shown extensive homology in more 3' regions; thus both mouse and primate repeats may derive from a common progenitor sequence (21). Furthermore, the presence of a common poly(A) rich 3' terminus to prime reverse transcription, and the occurrence of flanking short direct repeats (2) has led some investigators to propose that these sequences may propagate by retroposition (16). Surprisingly, long open reading frames have been identified in portions of both the primate (22,23) and mouse sequences (24) suggesting that these abundant interspersed repeats have potential protein coding functions.

In order to more fully understand the significance of these open reading frames, and to appreciate the sequence constraints during evolution, a mouse clone which hybridizes to both mouse BstNI 1.5kb and human HindIII 1.9kb restriction fragments was selected for further study; this clone contains a 5' region that has not been previously sequenced. Genomic blotting studies indicated this clone is fairly representative of other family members. We here show by sequence comparison that the human and mouse repeats are colinear in this 5' region, although important differences between each species are observed. The significance of the open reading frames, and other proposed biological "functions" are considered in relation to the high copy number and dispersal of these conserved LINES.

MATERIALS AND METHODS

Isolation and restriction mapping of clone containing BstNI 1.5kb related sequence

A partial EcoRI BALB/C genomic library cloned in charon 4A (obtained from Leslie Leinwand) was probed (25) with a nick-translated mouse genomic BstNI 1.5kb band which was isolated and electroeluted from a preparative agarose gel. Positive clones were counterscreened with a cloned nick-translated human HindIII 1.9kb fragment. A recombinant phage, here designated L7.1, selected with both probes, was mapped after cleavage with several restriction endonucleases; the resulting DNA digests were probed on Southern blots (26) with the BstNI 1.5kb mouse genomic probe and the HindIII 1.9kb human cloned fragments.

Plasmid subcloning

A EcoRI 5.8kb fragment released from the phage L7.1 was ligated into the EcoRI site of pUC8 (27). A recombinant plasmid, here designated L5.8, was characterized by restriction mapping and by Southern blotting to the nick-translated BstNI 1.5 kb genomic fragment.

DNA sequencing

For sequencing, L5.8 DNA was sonicated and repaired with S1 nuclease and Klenow polymerase (28). Sonicated fragments were fractionated by PEG precipitation (final concentration 7%) in 0.5 M NaCl. The average size of sonicated DNAs was about 500 bp; these fragments were ligated into the cloning/sequencing vector M13mp8 at the SmaI site. Phages containing BstNI-1.5 kb repeat were isolated and sequenced by the dideoxy chain

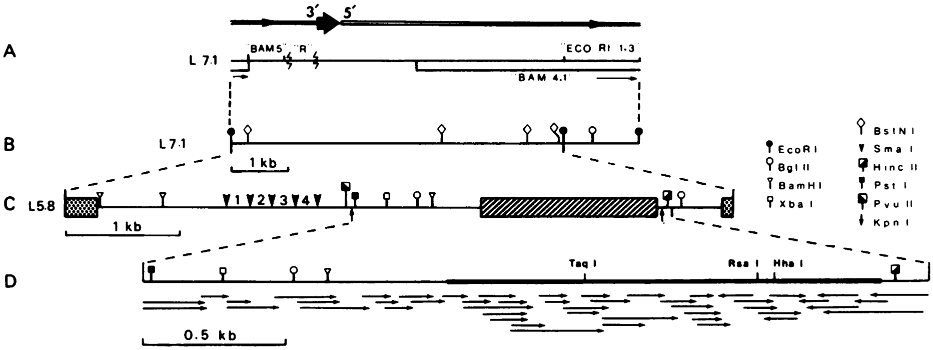

Fig 1: A) Definition of 3' and 5' regions of L7.1 with reference to previously described restriction fragments of the L1 repeat family. The "EcoRI 1.3" fragment represents a part of the "Bst 1.7 kb" repeat (3). Note the "Bam 4.1" repeat (open box) is seen at both ends of the clone, i.e. the clone contains an incomplete tandem L1 array. B) General restriction map of cloned insert in L7.1 and C) L5.8. L5.8 is a EcoRI cleaved subclone of L7.1. Regions of L5.8 that hybridized to the BstNI 1.5kb (▨) or BstNI 1.7 kb (▩) genomic fragments are shown. The latter are found at each end of the clone. D) Sequencing strategy: Each arrow represents the orientation and reading length of a discrete M13 subclone selected with the BstNI 1.5kb genomic fragment. The BstNI 1.5kb sequence was determined twice or more over 96.5% of its length; 68% of 5' flanking sequences were sequenced twice.

```
   1 GTACCTGCCTTGCAAGAAGAGAGCTTGCCTGCAGAGAATACTCTGCCCACTGAAACTAAGGAGAGTGCTA   70
  71 CCCTCCAGGTCTGCTCATAGAGGCTAACAGAGTCACCTGAAGAACAAGCTCTTAACAGTGACAACTAAAA  140
 141 CAGCTAGCTTCAGAGATTACCAGATGGCGAAAGGCAAACGTAAGAATCCTACTAACAGAAATCAAGACCA  210
 211 CTCACCATCATCAGAACGCAGCACTCCCACCCCACCTAGTCCTGGGCACCCCAACACAACCGXAAATCTA  280
 281 GACCCAGATTTAAAAACATTTCTCATGATGATGATAGAGGACATCAAGAAGGACTTTCATAAGTCACTTA  350
 351 AAGATTTACAGGAGAGCACTGCTAAAGAGTTACAGGCTCTTAAAGAAAAGCAGGAAAACACAGCCAAACA  420
 421 GGTGATGGAAATGAACAAATCCATACTAGAACTAAAAGGGGAAGTAGACACAATAAAGAAAACCCAAAGC  490
 491 GAGGCAACGTGGAGATAGAAACCCTAGGAAAGAGATCTGGAACCATAGATGCGAGCATCAGCAACAGAAT  560
 561 ACAAGAAATGGAAGAGAGAATCTCAGGTGCAGAAGATTCCATAGAGAACATCGACACAACAGTCAAAGAA  630
 631 AATACAAAATGCAAAAGGATCCTAACTCAAAACATCCAGGTAATCCAGGACACAATGAGAAGACCAAACC  700
 701 TACGGTTAATAGGAATTGATGAGAATGAXXXTTTTCAACTTAAAGGGCCAGCTAATATCTTCAACAAAAT  770
 771 AATAGAAGAAAACTTCCCAAACATAAAAAAAGAGATGCCCATGATCATACAAGAAGCATACAGAACTCCA  840
 841 AATAGACTGGACCAGAAAAGAAATTCCTCCCGACACATAATAATCAGAACAACAAATGCACTAAATAAAG  910
 911 ATAGAATATTAAAAGCAGTAAGGGAGAAAGGTCAAGTAACATATAAAGGAAGGCCTATCAGAATTACACC  980
 981 AGACTTTTCACCAGAGACTATGAAAGCCAGAAGAGCCTGGACAGATGTTATACAGACACTAAGAGAACAC 1050
                                        ▼
1051 AAATGCCAGCCCAGGCTACTATACCCAGCCAAACTCTCAATTACCATAGATGGAGAAACCAAAGTATTCC 1120
               BstN1                    AAGCT CATACGT  A       TA   TAC   A
                                        [
1121 ACGACAAAACCAAGTTCACACAATATCTTTCCACGAATCCAGCCCTTCAAAGGATAAT......AACAGA 1190
     CA     GCAA T C G G G T T G CA    C GG  T     AA  GA C CC GAAGGA G  CT

                                           AsnAsnGlnGluSerAsnHisSerThrAsnGl
1191 AAAGAAGCAATACAAGGACGGAAATCACGCCCT...AGAACAACCAAGAAAGTAATCATTCAACAAACCA 1260
        C T G  AGG  CA  C GT C  GC G  GCA A T  TG C A TT    AG CCATCG GG T

     nLysGlu  AspSerHisLysAsnArgMetProThrLeuThrThrLysIleLysGlySerAsnAsnTyrP
1261 AAAAGAA..GACAGCCACAAGAACAGAATGCCAACTCTAACAACAAAAATAAAAGGAGCAACAATTACT 1330
     GG     ACTG  T A  T ACCAGC  AATA C AGCT AC TC T   G T   ATCA   TCAC  A

     heSerLeuIleSerLeuAsnIleAsnGlyLeuAsnSerProIleLysArgHisArgLeuThrAspTrpLe
1331 TTTCCTTAATATCTCTTAATATCAATGGACTCAATTCCCCAATAAAAAGACATAGACTAACAGACTGGCT 1400
      AA AA  T  A  T AC  G A        A   G T     T        CG    GG  A T   A
                      ][
     uHisLysGlnAspProThrPheCysCysLeuGlnGluThrHisLeuArgGluLysAspArgHisTyrLeu
1401 ACACAAACAGGACCCAACATTCTGCTGCTTACAGGAAACCCATCTCAGGGAAAAAGACAGACACTACCTC 1470
      A A GT  A      T  G G     TA T                T TGC G     C   TAGG

     ArgValLysGlyTrpLysThrIlePheGlnAlaAsnGlyLeuLysLysGlnAlaGlyValAlaIleLeuI
1471 AGAGTGAAAGGCTGGAAAACAATTTTCCAAGCAAATGGACTGAAGAAACAAGCTGGAGTAGCCATTTTAA 1540
      A A A     AC TC  ..G  C      .        AAAC A   A G TA  G  T  A  CC  G

     leSerAspLysIleAspPheGlnProLysValIleLysLysAspLysGluGlyHisPheIleLeuIleLy
1541 TATCGGATAAAATCGACTTCCAACCCAAAGTTATCAAAAAAGACAAGGAGGGACACTTCATACTCATCAA 1610
      C  T       CA     TA    A C AAG       G      A  A  C  T A    A GG A

     sGlyLysIleLeuGlnGluGluLeuSerIleLeuAsnIleTyrAlaProAsnAlaArgAlaAlaThrPhe
1611 AGGTAAAATCCTCCAAGAGGAACTCTCAATTCTGAATATCTACGCACCAAATGCAAGGGCAGCCACATTC 1680
        ATC   T AA      A  G  AA T  C  A     A  T     C   A  G A   C   G

     IleArgAspThrLeuValLysLeuLysAlaTyrIleAlaProHisThrIleIleValGlyAspPheAsnT
1681 ATTAGAGACACTTTAGTAAAGCTCAAAGCATACATTGCACCTCACACAATAATAGTGGGAGACTTCAACA 1750
       A AGC AGTCC GAGTG C  AC  AG G  T A ACT C            A

     hrProLeuSerSerLysAspArgSerTrpLysGlnLysLeuAsnArgAspThrValLysLeuThrGluVa
1751 CACCACTTTCTTCAAAGGACAGATCGTGGAAACAGAAACTAAACAGGGACACAGTGAAACTAACAGAAGT 1820
      C T   G  AA  TTA  A     AACA G       A T    A   T  CCA G  T G ACTC  C

     lMetLysGlnMetAspLeuThrAspIleTyrArgThrPheTyrProLysThrLysGlyTyrThrPhePhe
1821 TATGAAACAAATGGACCTGACAGATATCTACAGAACATTTTATCCTAAAACAAAAGGATATACCTTCTTC 1890
      C  C C   GCA     . A   C           TC CC G  C   T   C  A AG   AAAAAAT

     SerAlaProHisGlyThrPheSerLysIleAspHisIleIleGlyHisLysThrGlyLeuAsnArgTyrL
1891 TCAGCACCTCACGGGACCTTCTCCAAAATTGACCATATAATTGGTCACAAAACAGGCCTCAATAGATACA 1960
             A   CAC    AT       C      C   C    AAGT   G TCT     GC A  GT
     ysAsnIleGluIleValProCysIleLeuSerAspHisHisGlyLeuArgLeuIlePheAsnLysAsnIl
1961 AAAATATTGAAATTGTCCCATGTATCCTATCAGACCACCATGGCCTAAGACTGATCTTCAATAAAAACAT 2030
        GA  A      A AA  AACTGT CC         AG  CAA C A   AGAAC   GGGTT  G A
                                          ][
     eAsnAsnGlyLysProThrPheThrTrpLysLeuAsnAsnThrLeuLeuAsnAspThrLeuValLysGlu
2031 AAATAATGGAAAGCCAACATTCACGTGGAAACTGAATAACACTCTTCTCAATGATACCTTGGTCAAGGAA 2100
      CTC C CA   C ACT  ACT  A        A  C   CTG  C  G     CTA  G   GC TA C

     GlyIleLysLysGluIleLysAspPheLeuGluPheAsnGluAsnGluAlaThrThrTyrProAsnLeuT
2101 GGAATAAAGAAAGAAATTAAAGACTTTTTAGAGTTTAATGAAAATGAAGCCACAACGTACCCAAACCTAT 2170
      A   G   GC      A  GATG  C  T  AACC G   G  CA   A      A    AGT T  C

     rpAspThrMetLysAlaPheLeuArgGlyLysLeuIleAlaLeuSerAlaSerLysLysLysArgGluTh
2171 GGGACACAATGAAAGCATTTCTAAGAGGGAAACTCATAGCGCTGAGTGCCTCCAAGAAGAAACGGGAGAC 2240
        ..   T C T   TG GTGT         T T     A  T A  T CA C  TGA  G T   A G

     rAlaHisThrSerSerLeuThrThrHisLeuLysAlaLeuGluLysLysGluAlaAsnSerProLysAr
2241 AGCACATACTAGCAGCTTGACAACACATCTAAAAGCCCTAGAAAAAAAGGAAGCAAATTCACCCAAGAG. 2310
     TTTCA A T GA  C C A   T    AT C     A        GC A .       A  TT    A C
            ][
     gSerArgArgGlnGluIleIleLysLeuArgGlyGluIleAsnGlnValGluThrArgArgThrIleGln
2311 GAGTAGACGGCAGGAAATAATCAAACTCAGGGGTGAAATCAACCAAGTGGAAACAAGAAGAACTATTCAA 2380
     T  C   A    A        T   A    A CA   C G  GG  A A  G   CA  A   CC

     ArgIleAsnGlnThrArgSerTrpPhePheGluLysIleAsnLysIleAspLysProLeuAlaArgLeuT
2381 AGAATTAACCAAACGAGGAGTTGGTTCTTTGAGAAAATCAACAAGATAGATAAACCCTTAGCTAGACTCA 2450
      A      TG  T C     C     T     A GG        A  T    G T AC    A     A

     hrLysGlyHisArgAspLysIleLeuIleAsnLysIleArgAsnGluLysGlyAspIleThrThrAspPr
2451 CTAAAGGGCACAGGGACAAAATCCTAATTAACAAAATCAGAAATGAAAAGGGAGACATAACAACAGATCC 2520
     TA  GAA A A CA  G  G CT  G  AG  TC   A A      T  A     T  C  C  C

     oGluGluIleGlnAsnThrIleArgSerPheTyrLys SerTyrThrGlnGlnAsnTrpLysThrTrpTh
2521 TGAAGAAATCCAAAACACCATCAGATCCTTCTACAAA.AGCTATACTCAACAAAACTGGAAAACCTGGAC 2590
     CAC      A    CT         GAA A       C C  C  TG   AT     A     T  A  A
                                                                    BstN1
     rLysTrpThrAsnPheTrpThrAspThrArgTyrGlnSerEnd
2591 GAAATGGACAAATTTCTGGACAGATACCAGGTACCAAAGTTGAATCAGGATCAAGTTGACCATCTAAACA 2660
             T     C  C    C    ACCC C    GAC A  CA    AG   C   ATC   G

2661 GTCCCATATCACCTAAAGAAATAGAAGCAGTTATTAATAGTCTCCCAACCAAAAAAAGCCCAGGACCAGA 2730
      A  A   A  GGCTCT     TATG   ACA  C     CT A           G  T

2731 TGGGTTTAGTGC  2742
        A  C CA
                ]
```

Fig 2: Sequence comparison of the mouse BstNI 1.5kb repeat with the human HindIII 1.9kb repeat. The mouse BstN 1.5kb sequence begins at position 1061 and ends at position 2588 (5' to 3'). Several nucleotides upstream of the BstNI 1.5kb region [denoted by X] could not be unequivocally determined. Underlined sequences were determined only on one strand, representing 3.5% of the total sequence shown. The HindIII 1.9kb repeat, represented below, shows only the bases which differ. The dots symbolize gaps introduced when both sequences are aligned (gap weight=5.0, length weight=0.3), and are included in the nucleotide count. The potential amino-acid translations of the longest open reading frame in the BstNI 1.5kb sequence are above the mouse sequence. The brackets delineate the domains defined in the text and in Fig 4.

termination procedure (29). The sequencing strategy is shown in Fig 1D. The repeat sequence was determined twice or more over 96.5% of the sequence.

Sequence Analysis

A data base was constructed by overlapping the random sequences isolated from the M13mp8 library (29). The final sequence was aligned with a human HindIII-1.9 kb representative (22) and examined by open reading frame analysis (31).

Genomic Blotting

Mouse liver genomic DNA was prepared (32), digested with restriction enzymes, fractionated on 0.9% agarose gels, and blotted on nitrocellulose or cross-linked (33) to Pall Biodyne Nylon membranes. DNA restriction fragments or subclones were labeled by nick-translation or primer extension with ^{32}P. Hybridizations were done at 65°C in 4 x SSC and filters were washed at 42°C in 0.15M NaCl, 0.015M Na citrate. Kodak X-Omat R film was exposed with an intensifying screen at -70°C for 24-72 hours.

RESULTS

Previous results have indicated that a 1.5kb BstNI band is visible over the background smear (3), which is consistent with the presence of many copies of this sequence. The abundance of this sequence was confirmed by screening positive colonies in a lambda genomic library with the 1.5kb probe; 1.6x10^{4} genomic copies of this repeat were so detected. The following describes salient features of this repeat.

Position of the BstNI 1.5kb segment in clones L7.1 and L5.8

The 7.1 kb cloned DNA is composed of two EcoRI fragments of 5.8 and 1.3 kb long. Since both the BstNI 1.5 kb and the Hind III 1.9 kb probes hybridized exclusively to the EcoRI 5.8 kb fragment, this fragment was subcloned (L5.8) for more detailed restriction analysis (Fig 1C). Digestion of L5.8 with BstNI revealed a single BstNI 1.5kb copy situated near the 5' terminus of the BamHI 4.1kb restriction fragment (7) of the L1 family as depicted in Fig 1A,C.

Sequencing of the BstNI-1.5 kb repeat

We have determined the complete nucleotide sequence of the BstNI 1.5kb repeat (Fig 2). The 1515 bp repeat begins at position 1061 and ends at position 2588. The base composition is typical of non-coding eukaryotic DNA, namely 42.3% A, 22.2% C, 16.4% G, and 19.1% T. This BstNI 1.5kb repeat does not seem to contain any complex internal repeats as revealed by dot matrix analysis (results not shown).

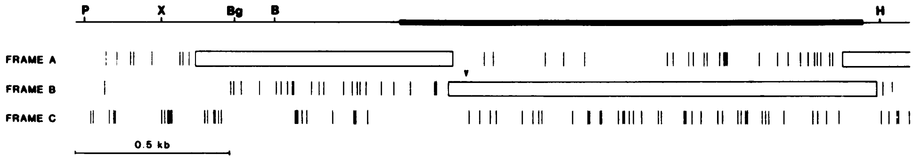

Fig 3: Distribution of the termination codons and location of the open reading frames in the 5'-3' orientation. Vertical lines indicate the position of the termination codons in each frame. The open reading frames (ORF) between two stop codons are represented as open boxes. The arrowhead shows the location of the first initiation codon (ATG) in frame b. The shaded box spans the length of the BstNI 1.5kb repeat. Only PstI (P), XbaI (X), BglII (Bg), BamHI (B) and HincII (H) sites are shown for clarity.

We also determined 1060bp of sequence 5' to the BstNI 1.5kb repeat. These results confirmed our previous restriction mapping and revealed numerous BstNI sites flanking the repeat at the following positions: 75, 251, 666, 675, 1016, and 2620; for simplicity not all of these are depicted in Fig 1C.

Open reading frame

The translation products in all three reading frames in the orientation 5' to 3' were examined. Fig 3 indicates that only two long open reading frames can be found. Any alternative reading frame has frequent stop codons. These two open reading frames overlap slightly and are frameshifted with respect to each other; the latter feature could be the result of a single base deletion in this family member, or both frames may be utilized as they are in other transposable elements (34). In frame b, the open reading frame extends 1401 bp (466 aa) and starts at position 1229. It is interesting to note that this long open reading frame is nearly as long as the repeat itself. In frame a, there is another potential coding sequence that extends 834 bp (280 aa) and starts at position 394. Since a long open reading frame was observed further downstream in the L1 family (24) it is possible that the reading frame continues in the 3' direction.

Sequence homology between mouse BstNI 1.5kb repeat and human Hind III 1.9kb repeat

The mouse BstNI 1.5kb and the human HindIII 1.9kb sequences were aligned using the Needleman-Wundsch algorithm (31). Allowing for a total of 13 gaps in the BstNI 1.5kb sequence at positions 1179 to 1184, 1224 to 1226, 1268, 1269, 2310 and 2558 and 7 gaps in the HindIII 1.9kb sequence at

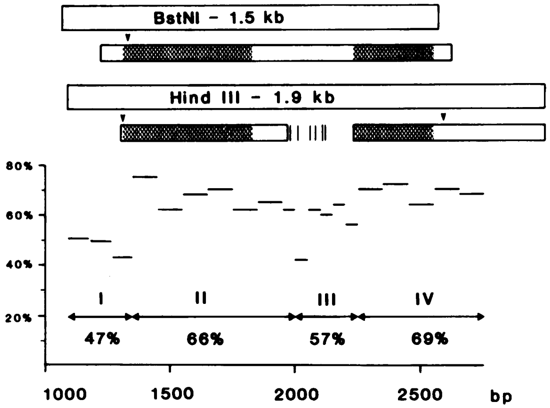

Fig 4: Homology comparison of members of the mouse BstNI 1.5kb and the human HindIII 1.9kb repeat families. The large open boxes illustrate the schematic alignment of the two families. The smaller open boxes below depict the open reading frames and the arrows represent the positions of initiation codons. The shaded boxes delimit the regions which show the most significant overall amino acid sequence homology. The ordinate of the graph indicates the percent of homology between the mouse and the human repeats as they were aligned along the abscissa in Fig 3. The alignment was separated in 4 domains based on the overall similarity of the two sequences.

positions 1489, 1490, 1501, 1839, 2174, 2175 and 2291, an alignment was found which spans the entire length of the mouse sequence starting at position 1086 (Fig 2). Although other matches were possible at lower stringency, the alignment in Fig 2 was clearly the most parsimonius, i.e. the fewest gaps and mismatches are introduced into either sequence in this pairing. The fact that these sequences are congruent throughout their respective lengths justifies the following analysis of their conservation by dividing them into domains. First, we divided the aligned sequences into segments (50bp, 80bp or 100bp depending upon the nature of the match) and defined 4 domains (I,II,III, and IV) based on the percent homology as shown in Fig 4 The first domain or 5' end of both sequences is not much better than alignment of random sequences having the same base composition. The most homologous regions between the mouse BstNI 1.5kb and human HindIII 1.9kb repeats are domains II and IV, the latter being the strongest one, with 69% homology. Domains II and IV correspond to the two open reading frames observed in the HindIII 1.9kb repeat.

B. Citron, et al (submitted) has immuno-precipitated two distinct human poly A$^+$ βDNA- galactosidase fusion proteins using a single antibody to pure phenylalanine hydroxylase. The coding sequences for these proteins contain overlapping regions from domain II. Since the mouse L1 sequence and now three independently derived human DNA clones share the same (and only possible) open reading frame, it is very likely the above assigned open reading frame (beginning at position 1229) is correct; a series of internal frameshifts can not account for these observations. Domain III is less homologous and contains a series of stop codons in all 3 forward frames in the human HindIII 1.9kb sequence. These may result from base substitutions in the HindIII 1.9kb sequence which interrupt the open reading frame. Although we did not sequence further than 154 bp downstream of the BstNI 1.5kb repeat, others have demonstrated that primate and rodent LINEs are similar for at least another 2kb in the 3' direction (21).

We tried to assess whether this conservation represents a selection for these frames across species boundaries, or was simply a coincidental sequence preservation. A significant protein homology was found within the two open reading frames in domains II (52.5%) and IV (58%) which parallels the similarity observed in the DNA sequence comparison. The conserved domains II and IV might be maintained simply at a DNA sequence level, or in addition, at a protein-coding sequence level. Fitch (35) has suggested that the number of nucleotide changes which result in amino-acid replacements, contrasted with those which preserve the coding potential in two homologous open reading frames, provides a measure of the conservation of the respective gene products. He found that the ratio of amino-acid replacements (R) to synonymous changes (S) is of the order 1.0 ± 0.3 for homologous genes, 1.7 ± 0.4 for pseudogenes and genes, and greater than 2.2 for weak or non homologous pairs. We used these criteria to ask whether domains II and IV were selected with respect to a potential L1 coding function. Table I shows that the value of R/S for each of the four domains correlate well with the known open reading frames present in mouse and human sequences. Since others have proposed that most members of the L1 family have characteristics of pseudogenes, we did not expect to find that the value of R/S for domain IV would correspond to that observed when two homologous genes are compared. This prompted us to determine base substitution rates (K_1, K_2, K_3) at each of the respective codon positions in order to estimate the evolutionary distance between the potential rodent and primate gene products (Table I, Ref. 36).

The values of K_1 and K_2 for domain IV are similar within standard errors, but K_3 is markedly higher, and this difference can be almost

Table 1. Conservation of the BstNI-1.5 kb and HindIII-1.9 kb Sequences at the Codon Level

Comparison	R/S	K_1	K_2	K_3	K_s
Domain I	>3				
Domain II	1.76	0.48±0.07	0.26±0.04	0.79±0.11	0.60±0.10
Domain III	2.15				
Domain IV	1.29	0.29±0.06	0.33±0.06	0.83±0.17	0.68±0.16
Human vs rat I preproinsulin A & B chains		0.04±0.03	0.00	0.46±12	0.38±0.12
Rabbit α - globin vs Rabbit β -globin		0.60±0.08	0.44±0.04	0.90±0.14	0.68±0.13
Human vs rat presomatotropins		0.26±0.04	0.18±0.03	0.53±0.07	0.44±0.07
Human β tubulin vs β tubulin processed pseudogene	1.36	0.03±0.01	0.04±0.01	0.06±0.02	0.05±0.02
Human metallothionine II vs processed pseudogene	0.25	0.02±0.02	0.00	0.07±0.06	0.07±0.06
Human immunoglobulin λ constant region vs processed pseudogene	.1.8	0.13±0.06	0.17±0.08	0.30±0.11	0.25±0.10

The values K_1, K_2, K_3 represent the number of base substitutions at the 3 codon positions as estimated using the model 3ST developed by Kimura (34). K_s denotes the synonymous change at position 3. For domains I and III, the calculations were not applicable. R/S is the number of non homologous codons which result in amino-acid replacements divided by those which translate into identical residues. (Sequences for comparison were derived from references 34, 37-39).

entirely attributed to synonymous codon changes (K_s). The frequency of base substitutions in positions 1 and 2 provide a measure for the rapidity of evolution in this L1 product region. In contrast with the rat and human insulin A and B peptides (36) for example, the L1 protein is quite dynamic. Relative to some processed retroposons, such as β tubulin and metallothionine (37,38), L1 "pseudogenes" are old. However, they have diverged less than the rabbit α and β globin genes, but more than rat and human presomatotropin (Table I).

The value of R/S for domain II (1.76) is characteristic of a mutated gene, and the observed base substitution rates at each of the coding positions bear this out. Like domain II, K_3 is larger than K_1 or K_2, but most of these third position changes are silent. However, K_1 is markedly higher than K_2, and changes at position 1 most likely explain the higher frequency of amino acid replacements in this domain. Sequences undergoing random drift such as tubulin do not show this disparity between K_1 and K_2 (Table I). It is possible that the conservation at codon position 2 in this domain may indicate a function that is unrelated to cytoplasmic protein coding (see Discussion).

Study of Genomic L1 Members

Genomic mapping studies indicated that the clone chosen for detailed evaluation here was highly representative of the most numerous genomic members by restriction analysis [Fig 6A and (10)]. We also detected a number of well-defined, albeit minor abundance fragments, which have not been previously mapped. A 1.4 kb MspI band was proposed to flank the major 3.6kb MspI band (10); we were able here to detect this band (Fig 6A,6B). In KpnI digests (Fig 6A) a strong 2.7kb band, and a weak 0.8kb band were also detected. Similar results were seen in long exposures with the 2.3kb BamHI probe mapped in Fig 6E (data not shown). Thus, most of the genomic L1 repeats conserve the 5' KpnI site (as seen in the clone), while a fraction of them also contain an additional KpnI site (0.8 kb downstream).

Differential genomic digests with HpaII and MspI were used to detect methylation within the genome at the 5' and 3' ends of L5.8. Our results indicate that, in adult liver DNA, all CpG sites recognized by these enzymes, including those which map to the 5' terminus of the majority of L1 members, are highly methylated.

Tandem arrangement of two L1 repeats in the lambda clone

The map in figure 1A shows the position of previously described fragments of this family as they are ordered in our selected lamda clone. Note part of the "BstNI 1.7" element is present at both ends of the clone. The left BstNI 1.7 member contains the 3' end of this element, and the right contains the 5' end, i.e. these segments were derived from two distinct L1 repeats. These two L1 repeats are tandemly arrayed in a head-to-tail fashion with no measurable intervening DNA. This is an unusual feature for an "interspersed" repeat. Given the high copy number of L1 repeats in the genome, the likelihood of randomly selecting a single unique clone with this tandem array would be less than 6 x 10^{-5}. It is therefore likely that other tandem L1 arrays are present in the genome, and blotting studies were consistent with this (vide infra). Presumably if the L7.1 insert was longer at the 3' end, it would continue through the rest of the "BstNI 1.7" -> "Bam5" -> "R" repeats, as seen at the left end of Fig 1A.

Tandem repeats upstream of the BstNI 1.5kb segment

Digestion of L5.8 with SmaI revealed 3 fragments of 3.6kb, 1.4kb and 200bp in length, the sum of which is less than the length of the total 5.8kb insert. We suspected that L5.8 carried several copies of the 200bp fragment, since this ethidium stained band was more intense than expected if it were present in a 1:1:1 stochiometry with the larger fragments. Partial digestions with SmaI revealed an evenly spaced ladder of restriction fragments separated by even lengths. This pattern was consistent with the presence of 4 tandemly arranged 200bp segments, each containing a SmaI site. The most downstream member of these tandem repeats is situated 1.3kb upstream from the BstNI 1.5 kb segment (Fig 1C). Blotting experiments with an M13mp8 subclone which spanned one repeat unit (clone 12.50) demonstrated that each of the ladder fragments were highly homologous (results not shown).

In order to 1) determine whether the linkage of SmaI subunits observed in L7.1 was a common feature of L1 genomic family members, and 2) further define the 5' terminus of the L1 family, we undertook a genomic blotting study using DNA probes derived from contiguous segments of L5.8. Hybridization with the 2.3kb BamHI probe (Fig 5B, lane Bm) revealed a ladder of tandem repeats spaced approximately 200bp apart. This ladder reflects an internally repeated BamHI restriction site in the SmaI array. Hybridization with the adjacent 2.3kb BglII fragment also highlighted this same ladder (Fig 5A, lane Bm). Thus, a large proportion of sequences upstream of the BstNI 1.5 repeat (including the SmaI array) are uniformly linked to L1 genomic family members. Since hybridized bands corresponding to SmaI units 1 and 2 were more intense than those of 3 and 4, genomic L1 members possessing 1 or 2 SmaI repeats may be more frequent. Alternatively, the SmaI units may be polymorphic with respect to the internal BamHI restriction site.

The SmaI repeats essentially define the 5' termini of most genomic L1 family members. Sequences of any significant length 5' to the SmaI units appear to be quite variable in genomic blots. For example, the 3' 2.3kb BglII probe yielded a clearer pattern of discrete genomic bands than the 5' 2.3kb BamH1 probe (Fig 5 A,B). Since blots probed specifically with the

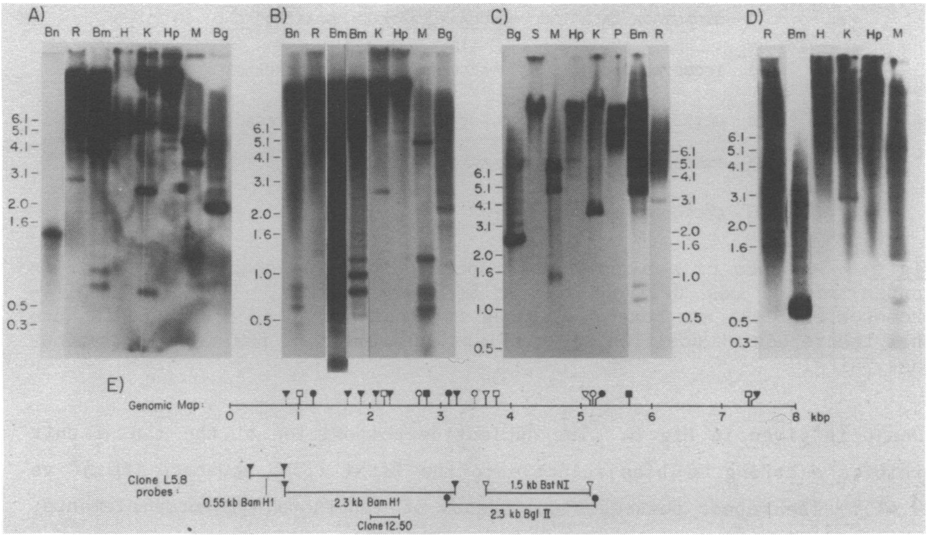

Fig 5: Genomic map. Genomic mouse liver DNA was digested with Bam HI (B), BstNI (Bn), KpnI (K), MspI (M), HpaII (Hp), HindIII (H), BglII (Bg), EcoRI (R) and PstI (P) endonucleases. Aliquots (10 ug) were fractionated on a 0.9% agarose gel and stained with ethidium bromide. Calibration in kb. DNAs were transferrred to Biodyne nylon membranes and probed with the following fragments from L5.8 : A= 2.3 kb BglII probe, except lane Bn probed with the BstNI 1.5kb fragment; B= 2.3 kb BamHI probe; C= clone 12.50; D= 0.55 kb BamHI probe.
E: Interpretation of genomic blotting data. Conserved restriction sites in the genomic map are represented with solid stems, whereas less abundant sites have dotted stems. BamHI (▼), BglII (●), KpnI (o), MspI (□), BstNI (▽), EcoRI (■).

SmaI repeat (clone 12.50) did not show this background smear (Fig 5C), it is unlikely that this region of the L1 family is extensively rearranged.

In long autoradiographic exposures, it was possible to detect a set of minor, constant-length bands extending upstream from the SmaI repeats (Fig 6D, lane Bm). These bands also hybridize to a probe corresponding to the "Bam5" region (Fig 1A), which is normally found at the 3' end of most L1 repeats. Thus, there are detectable, but infrequent tandem (head to tail) L1 repeats in the genome.

Notable sequence features of the SmaI repeat unit

In view of the proximity (approximately 600 bp) of the SmaI repeats to the first putative open reading frame (frame A), it was of interest to ask if any unusual sequence features, i.e. potential regulatory singnals were present in the SmaI repeats. The sequence of a single representative SmaI

```
  1   GTGAGTGGATCACAGTGCCTGCCCCAATCCAATCGCGCGGAACTCGAGAC   50

 51   TGCGGTACATAGGGAAGCAGGCTACCCGGGCCTGATCTGGGGCACAAGTC  100

101   CCTTCCGCTCGACTCGAGACTCGAGCCCCGGGTACCTTGACAGCAGAGTC  150

151   TTGCCCAACACCCGCATGGCCCAAAAGGGGACTCCCCACGGGACCCTAAG  200

201   CCCTCTGG  208
```

Fig 6: Sequence of a single SmaI repeat unit. The sequence of a single tandem repeat was derived from M13mp8 recombinant, 12.50. Putative promoter elements are boxed. A 10 bp internal repeat is underlined. Note that the repeat at position 119 overlaps the repeat at position 112 by four base pairs.

repeat is given in Fig 6. The nucleotide composition of the tandem unit exhibits a strong GC bias relative to the BstNI 1.5kb sequence (60.5% vs 39.4%). The repeat possesses components of functional promoter elements, i.e. a duplicated CCAAT box at position 25, separated from an ATA signal at position 59. These sequences were observed in other repeat members at equivalent positions. A 10 base sequence (ACTCGAGACT), found between these elements, recurs at positions 112 and 119. This latter feature is analogous to the cluster of repeated oligonucleotides seen in the vicinity of the β-like globin, herpes thymidine kinase and SV40 early promoters.

DISCUSSION

We have isolated a 7.1kb clone from a mouse genomic lambda library which fairly represents L1 repeats in the genome. Two BstNI 1.7kb elements were found at each end of this clone, consistent with a head to tail arrangement of two L1 repeats. A limited number of L1 repeats are also tandemized in the genome; this arrangement was unexpected for a sequence with an "interspersed" or even a clustered chromosomal distribution (40,41). Several LINE elements with rearrangements that are not characteristic of transposition or homologous recombination events have been reported (23). Some of the rearrangements might be produced by unequal crossover (42) between two different repeats, and such a mechanism could lead to the tandemization observed here.

The new mouse sequence data here (2.5kb from the 5' region of L1) allowed comparison of mouse and corresponding human sequences. This data confirmed our previous hybridization results (17) which had been questioned (43), and indicated significant (up to 70%) sequence homology. Singer, et. al (21) have been able to detect comparable sequence homologies in the adjacent 3' end of the mouse L1 and the primate KpnI families. This new data indicates extremely long and contiguous sequence homology in the LINES of widely divergent species.

A retroposon hypothesis has been put forward to explain the mechanism of proliferation of L1 elements (16). Retroposons often arise from sequences having open reading frames (as for example, processed pseudogenes), but may not necessarily maintain coding potential. The sequence decay of non-functional retroposons often parallels that seen in pseudogenes. The primate-rodent sequence comparison revealed that the two regions (domains II and IV) of strongest sequence homology coincided precisely with two open reading frames. Alternating regions of weaker homology in the human homolog e.g. domain III are excluded from coding for a gene product because of translational stop codons in all three frames (21). Our sequence analysis of domain IV suggests that the coding potential of a segment of the L1 family has been retained over 120 million years. Furthermore, two cDNAs which include domain II have been identified, and these appear to use the same phase as depicted here (see Results, B. Citron, et al submitted). Thus this region of the L1 family remains under active selection, possibly encoding a conserved gene product present throughout the mammalian radiation.

Although L1 repeats may have originated from a protein coding sequence, additional other functions unrelated to protein products may have put constaints on sequence divergence in this family; such other functions could have evolved concommitant with the advent of high copy numbers of this element in the genome. These numerous copies have been classified as pseudogenes (44) because infrequent stop codons separate long open reading frames in these sequences. The bias in base substitutions observed at each of the 3 codon positions in domain II is not normally seen in other processed pseudogenes (Table I). Indeed, the term pseudogene may loosely describe some sequences whose functions are not yet fully appreciated. With respect to non-coding or possibly pleiotropic functions, the following need to be considered: 1) Previous in-situ hybridization studies have indicated that the related human HindIII 1.9kb repeats are generally "clustered" in Giemsa-dark banding regions, where they may play a role in chromosome alignment, looping, or site recognition (16,41,45). If the mouse L1 repeats are similarly organized on chromosome arms, this might suggest a common periodicity and conserved structural role for this set of mammalian interspersed repeats. Specific proteins might participate in this periodic structure, and these repeats have been associated with the nuclear matrix proteins (46). 2) These sequences could have a regulatory role, possibly via RNA intermediates (47). Although these LINES can be abundantly represented in hnRNA, transcripts are inefficiently exported to the cytoplasm and translated in Hela and erythroid cells (13,15). It is conceivable however, that these transcripts bind proteins in the nucleus, and act at this site. Transcribed molecules of this repeat are present in sufficient concentrations to titrate specific nuclear constituents [in an analogous manner to 5S RNA (48)]. The preferred conservation of position 2 in most codons (in domain II) could provide the sequence specificity necessary to stabilize such interactions. 3) Pleiotropic coding and regulatory roles for DNA sequences are not unprecedented in both procaryotes and eukaryotes (49-51). It is therefore also conceivable that the observed pattern of nucleotide conservation in domain II could represent an underlying regulatory element in an apparent open reading frame.

A short tandem repeat, designated here as SmaI, was found between the two L1 repeats in clone L7.1. This SmaI unit is also highly conserved in the genome and linked to the BstNI 1.5kb repeat in many members of the L1 family. Data here indicated that this SmaI region may generally define the 5' end, or linking region, in most L1 family members. We do not know if SmaI repeats may also reside at other genomic sites, separated from the L1 family. By serving as a recombinational target, the SmaI repeat could have had an important role in the generation of adjacent arrays of L1 elements. However, since 5' L1 elements including the SmaI repeats are less abundant than 3' elements (2), the genomic association of two or more full-length L1 family members may be somewhat infrequent.

To ensure their own propagation, the L1 repeats, if they are retroposons, may atypically carry their own promoter and/or enhancer sequences. Indeed, the 5' SmaI tandem repeats appeared to contain a promoter sequence. Although sequence repetition is not an essential feature of enhancer elements, direct tandem repeats can potentiate transcriptional activity under selective conditions non-permissive for a truncated enhancer (52,53). The genomic ladder of SmaI repeats we have observed could be consistent with the presence of L1 members possessing different frequencies of 5' terminal repetition. Retroposons, when they are reinserted into the genome, usually do not carry their single regulatory sequences. Thus they are transcriptionally incapacitated. However, if L1 repeats contain two or more SmaI "regulatory" units, a retroposed L1 element could carry its second SmaI unit, which is necessary for viable transcription, i.e. ensuring transcription via a "recursive regulatory" function. Further functional analyses of this SmaI element, in experiments where it may be linked to a test gene, can be used to determine if it has any effect on expression.

ACKNOWLEDGEMENTS

We are grateful to T. Fanning for the gift of the Bam5 clone, H. Zachau for the RI/RII clone, S. Weissman for reading the manuscript, and R. Chesney for patient typing. This work was supported by NIH grant CA 15044

REFERENCES

1. Lewin, B. (1980) Gene Expression II-Eukaryotic Chromosomes, pp. 503-530, John Wiley & Sons, New York.
2. Voliva, C.F., Jahn, C.L., Comer, M.B., Hutchison, C.A. III, and Edgell, M.H. (1983) Nucleic Acids Res. 11, 8847-8859.
3. Manuelidis, L. (1980) Nucleic Acids Res. 8, 3247-3258.
4. Chen, S. and Schildkraut, C. (1980) Nucleic Acids Res. 8, 4075-4090.
5. Heller, R. and Arnheim, N. (1980) Nucleic Acids Res. 8, 5031-5041.
6. Brown, S.D.M. and Dover, G.A. (1981) J. Mol. Biol. 150, 441-466.
7. Meunier-Rotival, M., Soriano, P., Cuny, G., Strauss, F., and Bernardi, G. (1982) Proc. Natl. Acad. Sci. 79, 355-359.
8. Fanning, T.G. (1982) Nucleic Acids Res. 10, 5003-5013.
9. Gebhard, W. and Zachau, H.G. (1983) J. Mol. Biol. 170, 255-270.
10. Meunier-Rotival, M. and Bernardi, G. (1984) Nucleic Acids Res. 12, 1593-1608.
11. Heller, D., Jackson, M., and Leinwand, L. (1984) J. Mol. Biol. 173, 419-436.
12. Shafit-Zagardo, B., Brown, F.L., Zavodny, P.J., and Maio, J.J. (1983) Nature 304, 277-280.
13. Kole, L.B., Haynes, S.R., and Jelinek, W.R. (1983) J. Mol. Biol. 165, 257-286.
14. Lerman, M.I., Thayer, R.E., and Singer, M.F. (1983) Proc. Natl. Acad. Sci. 80, 3966-3970.
15. Schmeckpeper, B., Scott, A., Smith, L. J. Biol. Chem., 259, 1218-25 (1984).
16. Fanning, T.G. (1983) Nucleic Acids Res. 11, 5073-5091.
17. Manuelidis, L. (1982) In: "Genome Evolution and Phenotypic Variation" G.A. Dover and R.B. Flavell, eds., pp 263-285. Academic Press, New York.
18. Manuelidis, L. and Biro, P.A. (1982) Nucleic Acids Res. 10, 3221-3239.
19. Singer, M.F. (1982) Int. Rev. Cytol. 76, 67- 112.
20. Adams, J., Kaufman, R.E., Dretschner, P.J., Harrison, M., and Nienhuis A.W. (1980) Nucleic Acids Res. 8, 6113-6129.
21. Singer, M.F. Thayer, R.E., Grimaldi, G., Lerman, M.I., and Fanning T.G. (1983) Nucleic Acids Res. 11, 5739-5745.
22. Manuelidis, L. (1982) Nucleic Acids Res. 10, 3211-3219.
23. Potter, S.S. (1984) Proc. Natl. Acad. Sci. 81, 1012-1016.
24. Martin, S.L., Voliva, C.F., Burton, F.H., Edgell, M.H., and Hutchison, C.A. III (1984) Proc Natl Acad Sci 81, 2308-2312.

25. Benton, W.D. and Davis, R.W. (1977) Science 196, 180-182.
26. Southern, E.M. (1975) J. Mol. Biol. 38, 303-517.
27. Messing, J. and Viera J. (1982) Gene 19, 269-276.
28. Deininger, P. (1983) Analytical Biochemistry 129, 216-223.
29. Sanger, F., Nicklen, S., Coulsen, A.R. (1977) Proc. Natl. Acad. Sci. 74,5463-5467.
30. Staden, R. (1980) Nucleic Acids Res. 8, 3673-3694.
31. Devereux, J., Haeberli, P., and Smithies, O. (1984) Nucleic Acids Res. 12, 387-395.
32. Maniatis, T., Fritsch, E., and Sambrook, J. (1982) Molecular cloning, Cold Spring Harbor Laboratory, 280-282.
33. Church, G.M. and Gilbert W. (1984) Proc. Natl. Acad. Sci. 81, 1991-1995.
34. Jacks, T. and Varmus, H.E. (1985) 230, 1237-1241.
35. Fitch, W.M. (1980) J. Mol. Evol. 16, 153-209.
36. Kimura, M. (1981) Proc. Natl. Acad. Sci. 78, 454-458.
37. Wilde, C.D., Crowther, C.E., Cripe, T.P., Lee, M.S., and Cowan, N.J. (1982) Nature 297, 83-84.
38. Karin, M. and Richards, R.I. (1982) Nature 299, 797-802.
39. Hollis, G.F., Hieter, R.A., McBride, O.W., Swan, D., and Leder, P. (1982) Nature 296, 321-323.
40. Shafit-Zagardo, B., Maio, J.J., Brown, F.L. (1982) Nucleic Acids Res. 10, 3175-3179.
41. Manuelidis, L. and Ward, D.C. (1984) Chromosoma 91, 28-38.
42. Jones, R.S. and Potter, S.S. (1985) Proc. Natl. Acad. Sci. 82, 1989-1993.
43. Grimaldi, G. and Singer, M.F. (1983) Nucleic Acids Res. 11, 321-328.
44. Martin, S.L., Voliva, C.F., Burton, F.H., Edgell, M.H., Hutchison, C.A.III, (1984) Proc. Nat. Acad. Sci. 81, 2308-12.
45. Small, D., Nelkin, B. and Vogelstein, B. (1982) Proc. Natl. Acad. Sci. 79, 5911-5915.
46. Chimera, J.A. and Musich, P.R. (1985) J. Biol. Chem. 260, 9373-9.
47. Kramer, A., Keller, W., Appel, B., and Luhrmann, R. (1984) Cell 38, 299-307.
48. Pelham,H.R.B. and Brown, D.D. (1980) Proc. Natl. Acad. Sci. 77, 4170-4174.
49. Fisher, R. and Yanofsky, C. (1984) Nucleic Acid Res. 12, 3295-3302.
50. Landick, R. and Yanofsky, C. (1984) J. Biol. Chem. 259,11550-11555.
51. Wright, S., Rosenthal, A., Flavell, R. and Grosveld, F. (1984) Cell 38, 265-273.
52. de Villers, J., Schaffner, W., Tyndall, C., Lupon, S., Kamen, R. (1984) Nature 312, 242-246.
53. Weber, F., de Villiers, J., and Schaffner, W. (1983) Cell 36, 983-992.