Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 2308–2312, April 1984
Biochemistry

A large interspersed repeat found in mouse DNA contains a long open reading frame that evolves as if it encodes a protein

(*L1* family/*Bam*HI family/*Kpn* I family/DNA sequence/gene correction)

SANDRA L. MARTIN*, CHARLES F. VOLIVA*†, FRANK H. BURTON*, MARSHALL H. EDGELL*†‡, AND CLYDE A. HUTCHISON III*†‡

*Department of Microbiology and Immunology, †Curriculum in Genetics, and ‡Program in Molecular Biology and Biotechnology, 804 Faculty Laboratory and Office Building, University of North Carolina, Chapel Hill, NC 27514

Communicated by Robert L. Sinsheimer, December 27, 1983

ABSTRACT **DNA sequence analysis of a region contained within a large, interspersed repetitive family of mice reveals a long open reading frame. This sequence extends 978 base pairs between two stop codons, creating a reading frame that is open for 326 amino acids. The DNA sequence in this region is conserved between three distantly related *Mus* species, as well as between mouse and monkey, in a manner that is characteristic of regions undergoing selection for protein function.**

Interspersed repetitive elements are ubiquitous in the DNA of higher eukaryotes. They are found in organisms as diverse as frogs (1), sea urchins (2), flies (3), birds (4), rodents (5), and humans (6); nevertheless, the reason for their existence is not understood. Hypotheses concerning their contribution to the normal patterns of gene expression in differentiated cells range from a model in which they are seen as the main perpetrators of regulation (7, 8) to the extreme opposite view of them as neutral hitchhikers in the genome (9, 10). Other hypotheses suggest they behave as transposable elements. As transposons, repetitive families may create new patterns of expression by providing new promoters or physically disrupting gene function following an insertion (11), or they could provide a constant source of new material for evolution by keeping the genome in a constant state of flux through homologous recombination events (12).

It has long been hoped that studies of the structure and organization of interspersed repetitive families would suggest possible functions for such sequences. Structural investigations have revealed two basic types of interspersed repetitive families, short and long (reviewed in ref. 13). Members of the short families tend to be <500 base pairs (bp) long and are represented upwards of 10^5 times in the genome. RNA copies of these elements are found in cells, and *in vitro* their cloned DNA is transcribed by RNA polymerase III. Prominent members of this class include the *Alu* family of primates and the *B1* family of rodents (see ref. 14 for review). Less is known about the longer types of repetitive elements than the short because they have been studied in detail only recently. Members of these repeat families can be >5 kilobases (kb) in length and represented up to 10^5 times in the genome (reviewed in ref. 13).

In mice, one of the major families of long interspersed repeats is the *L1* family§ (15), which has also been called the *Bam*HI or *MIF-1* family (16, 17). *L1Md* family members range up to 7 kb in length (17) and are homologous to the primate *L1* family, previously called the *Kpn* I family (18, 19), which has also been characterized in some detail (see ref. 13 and references therein). Transcripts from portions of these large repeats have been detected in both mice (11, 16, 20) and primates (19, 21–23). There is some evidence that these sequences are associated with polyribosomes in mouse liver (16), although they were not detected in polyribosomal preparations from human culture cells (21). The presence of the sequences in polysomes does not necessarily imply protein coding function because *Alu* family transcripts have also been detected in polysome preparations (see ref. 24 for review).

In collecting DNA sequence data on various members of the *L1* family for other purposes, we were struck by the presence of a long open reading frame in part of the repeat family sequence. This open reading frame is noteworthy because it coincides with a region of the repeat that is represented in mRNA (16, 20) and is conserved among three distantly related species of *Mus* (*domesticus*, *caroli*, and *platythrix*) as well as between mouse and monkey. These observations, along with the properties of the sequence of the open reading frame, suggest that this portion of the *L1* repeat family has been functional for most, if not all, of its evolutionary history.

MATERIALS AND METHODS

Restriction endonucleases, DNA polymerase large fragment, T4 polynucleotide kinase, and 26-bp primer were purchased from New England BioLabs or Bethesda Research Laboratories. [γ-^{32}P]ATP and [α-^{32}P]dATP were purchased from New England Nuclear. Liver from *M. platythrix* was obtained through Michael Potter from Litton Bionetics. High molecular weight DNA was prepared essentially as described by Kan and Dozy (25). Genomic DNA from *M. domesticus* and *M. caroli* was the gift of R. Padgett. Clones containing *L1Md*-4 were isolated originally from a genomic library prepared from BALB/c sperm DNA in Charon 4A (26).

Random isolates of the *Bam5* (500-bp *Bam*HI fragment) subset of the *L1* repeat were cloned into M13mp7. Genomic DNA from *M. domesticus*, *M. caroli*, and *M. platythrix* was digested with the restriction endonuclease *Bam*HI. After electrophoresis through 1% agarose, the 500-bp ethidium staining band was recovered by electroelution and then ligated into the *Bam*HI site of M13mp7. About 10% of the transformants hybridized to the single-stranded, radiolabeled *Bam5* probe. Under the conditions used, sequences sharing at least 65–70% homology with the probe would be detected. It is not likely, however, that many of the *L1* family se-

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.

Abbreviations: URF, unidentified reading frame; kb, kilobase(s); bp, base pair(s).

§The name *L1*, or *LINE 1*, has recently been suggested (15) as a replacement for the names previously used. The family name is followed by a two-letter genus and species designation, such as *L1Md* for the *L1* family in *Mus domesticus*. This new name overcomes the confusion inherent when different restriction endonuclease names are used to name homologous repeat families in different species.

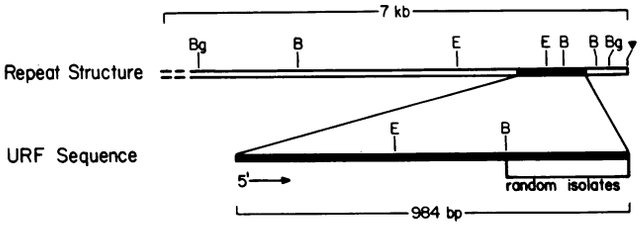

FIG. 1. Basic structure of *L1Md* family members and location of the open reading frame. The structure of a long family member is shown by the open bar. The filled in region indicates the open reading frame. Dashes are used at the 5′ end of the repeat structure because it is not known precisely where the end point is. The 316-bp subset of the URF that was subjected to sequence analysis in the random isolates is indicated. Some of the common sites for cleavage by restriction endonucleases are shown: *Bgl* I, Bg; *Bam*HI, B; *Eco*RI, E. The ▼ marks the A-rich 3′ end of the repeat. Contained within this structure are several subfamilies (usually restriction fragments) that are highly repetitive and have been described by others. These include the 1.3-kb *Eco*RI fragment (29–31), the 500-bp (20) and 4.0-kb (32) *Bam*HI fragments, and the region containing the 3′-most end of the repeat structure (33) that is linked to *Bam*5 (34). The truncated nature of this family and the degree of repetitiveness of the various regions are documented elsewhere (15).

quences that have diverged >80% from the probe still retain the two *Bam*HI sites defining the *Bam*5 fragment. We estimate that somewhat less than half of the sequences homologous to this region of the repeat actually have the *Bam*5 fragment, or roughly 20,000 copies per genome.

The DNA sequence of *L1Md*-4 was determined by using a combination of the chemical degradation (27) and the dideoxy chain elongation (28) methods. Sequences of the random isolates of *Bam*5 were determined by the chain elongation method.

RESULTS AND DISCUSSION

Location and Extent of the Unidentified Reading Frame (URF) in the *L1* Repeat Family. The dispersed, highly repetitive family of DNA sequences in the mouse, called *L1Md*, has an unusual structure. The 3′ end of the structure is conserved among copies, whereas the 5′ end is truncated in most family members at apparently random distances from the conserved end point (15, 17). This 3′ end is defined by an A-rich region, suggesting a role for RNA intermediates in the dispersal of the family. The structure of a long member of this family is shown in Fig. 1. In addition to being dispersed throughout the genome, portions of this repeat are found at seven locations within the β-globin gene cluster of the BALB/c mouse (15).

The DNA sequence from a portion (Fig. 1) of one of these repeats, *L1Md*-4, located between the $\beta h3$ and $\beta 1^{dmaj}$ genes, has been determined (Fig. 2). Within this sequence there is an open reading frame that extends 978 bp between two stop codons, TGA and TAG; the longest distance between a possible initiation codon and a terminator is 864 bp (Fig. 2). In examining the sequences that are immediately adjacent to these for the canonical sequences thought to be important in signaling precise and efficient transcription (36), nothing obvious has been found. Thus, if this is a functional unit, we are unable to predict at this time whether the entire structure is contained in this region or if this represents just a portion of a larger, perhaps multipartite, gene.

Is This URF Open Due to Chance? The preservation of this open reading frame seems unlikely to be due purely to chance. In a random stretch of 978 bp, the probability that one of the six possible reading frames will be open is 9.5×10^{-7} (Fig. 3).

Furthermore, this URF is found in random isolates of a subfragment from the *L1* repeat in three species. The DNA sequence of a subset of this open reading frame has been

```
                                                    50                                                  100
*** Ile Asp Pro Tyr Leu Ser Pro Cys Thr Lys Val Lys Ser Lys Trp Ile Lys Glu Leu His Ile Lys Pro Glu Thr Leu Lys Leu Ile Glu Glu Lys Val
TGAATCGATCCATACTTATCTCCTTGTACTAAGGTCAAATCTAAGTGGATCAAGGAACTTCACATAAAACCAGAGACACTGAAACTTATAGAGGAGAAAGTG

                                                   150                                                  200
Gly Lys Ser Leu Glu Asp Met Gly Thr Gly Glu Lys Phe Leu Asn Arg Thr Ala Met Ala Cys Ala Val Arg Ser Arg Ile Asp Lys Trp Asp Leu Met Lys
GGGAAAAGCCTTGAAGATATGGGCACAGGGGAAAAATTCCTGAACAGAACAGCAATGGCTTGTGCTGTAAGATCGAGAATTGACAAATGGGACCTAATGAAA

                                                   250                                                  300
Leu Gln Ser Phe Cys Lys Ala Lys Asp Thr Val Tyr Lys Thr Lys Arg Pro Pro Thr Asp Trp Glu Arg Ile Phe Thr Tyr Pro Lys Ser Asp Arg Gly Leu
CTCCAAAGTTTCTGCAAGGCAAAAGACACTGTCTATAAGACAAAAAGACCACCAACAGACTGGGAAAGGATCTTTACCTATCCTAAATCAGATAGGGGACTA

                                                   350                                                  400
Ile Ser Asn Ile Tyr Lys Glu Leu Lys Lys Val Asp Leu Arg Lys Ser Asn Asn Pro Leu Lys Lys Trp Gly Ser Glu Leu Asn Lys Glu Phe Ser Pro Glu
ATATCCAACATATATAAAGAACTCAAGAAGGTGGACCTCAGAAAATCAAATAACCCCCTTAAAAAATGGGGCTCAGAACTGAACAAAGAATTCTCACCTGAG
                                                                                                 Eco RI
                                                   450                                                  500
Glu Tyr Arg Met Ala Glu Lys His Leu Lys Lys Cys Ser Thr Ser Leu Ile Ile Arg Glu Met Gln Ile Lys Thr Thr Leu Arg Phe His Leu Thr Pro Val
GAATACCGAATGGCAGAGAAGCACCTGAAAAAATGTTCAACATCCTTAATCATCAGGGAAATGCAAATCAAAACAACCCTGAGATTCCACCTCACACCAGTC

                                                   550                                                  600
Arg Met Ala Lys Ile Lys Asn Ser Gly Asp Ser Arg Cys Trp Arg Gly Cys Gly Glu Arg Gly Thr Leu Leu His Cys Trp Trp Asp Cys Arg Leu Val Gln
AGAATGGCTAAGATCAAAAATTCAGGTGACAGCAGATGCTGGCGAGGATGTGGAGAAAGAGGAACACTCCTCCATTGTTGGTGGGATTGCAGGCTTGTACAA

                                                   650                                                  700
Pro Leu Trp Lys Ser Val Trp Arg Phe Leu Arg Lys Leu Asp Ile Val Leu Pro Glu Asp Pro Ala Ile Pro Leu Leu Gly Ile Tyr Pro Glu Glu Ala Pro
CCACTCTGGAAATCAGTCTGGCGGTTCCTCAGAAAATTGGACATAGTACTACCGGAGGATCCAGCAATACCTCTCCTGGGCATATATCCAGAAGAAGCCCCA
                                                                  Bam HI
                                                   750                                                  800
Thr Gly Lys Lys Asp Thr Cys Ser Thr Met Phe Ile Ala Ala Leu Phe Ile Ile Ala Arg Asn Trp Lys Glu Pro Arg Cys Pro Ser Thr Glu Glu Trp Ile
ACTGGTAAGAAGGACACATGCTCCACTATGTTCATAGCAGCCTTATTTATAATAGCCAGAAACTGGAAAGAACCCAGATGCCCCTCAACAGAGGAATGGATA

                                                   850                                                  900
Gln Lys Met Trp Tyr Ile Tyr Thr Met Glu Tyr Tyr Ser Ala Ile Lys Lys Asn Glu Phe Met Lys Phe Leu Ala Lys Trp Met Asp Leu Glu Ser Ile Ile
CAGAAAATGTGGTACATCTACACAATGGAGTACTACTCAGCTATTAAAAAGAATGAATTTATGAAATTCCTAGCCAAATGGATGGACCTGGAGAGCATCATC

                                                   950                                                  984
Leu Ser Glu Val Thr Gln Ser Gln Arg Asn Ser His Asn Met Tyr Ser Leu Ile Ser Gly Tyr ***
CTGAGTGAGGTAACACAATCACAAAGGAACTCACACAATATGTACTCACTGATAAGTGGATACTAG
```

FIG. 2. DNA sequence of the URF in *L1Md*-4 and its translation product. The *Eco*RI and *Bam*HI cleavage sites as well as the first methionine codon are underlined. Over 90% of the sequence has been confirmed in multiple experiments, and most of it has been determined on both strands. The experimental details will be published elsewhere. The deduced amino acid sequence was compared to sequences in the Dayhoff protein sequence data base (35) by using the program SEARCH. No sequences with significant homology to *L1Md*-4 were found.

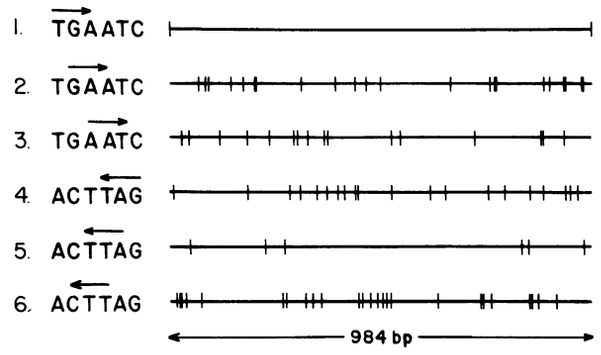

FIG. 3. Location of termination codons in six possible reading frames. Vertical hatch marks indicate the position of TAA, TGA, or TAG in each frame. The first six nucleotides of the *L1Md*-4 sequence are shown on the left; the arrow indicates the position and direction of the frame diagramed alongside. Four of the six possible frames contain no open regions of significant length (frames 2, 3, 4, and 6). Frame 5, which shares third positions with frame 1, has relatively fewer termination codons. This can be explained by constraints operating on frame 1. If the codon usage in frame 1 is considered, the number of termination codons observed in frame 5 does not differ from expectation.

determined for several copies of the repeat that were selected at random from the genomes of *M. domesticus* (BALB/c), *M. caroli*, and *M. platythrix*. The sequence of 10 clones containing the 500-bp *Bam*HI fragment from each species was determined for 315 bp between the *Bam*HI site and the terminator (see Fig. 1 for its position relative to the structure of the entire repeat). The consensus sequences derived for each of the three *Mus* species, *domesticus*, *caroli*, and *platythrix*, contain an open reading frame (Fig. 4) that corresponds to the one found in *L1Md*-4. In the 12 isolates from BALB/c (two from the globin region, plus 10 isolated from random locations in the genome), 11 have an open reading frame over these 104 amino acids. The single *Bam*5 isolate that does not have an open reading frame has suffered two single base-pair deletions; the resulting frameshift causes termination codons to come into phase further downstream. This is the most divergent representive of the repeat family that was isolated from BALB/c. Another random isolate of the *Bam*5 fragment from *M. domesticus* (the inbred line GR/A) has been subjected to sequence analysis and found to have terminators (20). In *M. caroli*, 6 of 10 random isolates have retained their open reading frames. Of the 4 with termination codons, 2 result from point substitutions and 2 from deletions that create a frameshift. Nine of the 10 sequences from *M. platythrix* also share the open reading frame. The tenth has a base substitution that results in a termination codon.

Another aspect of these DNA sequence data from the three *Mus* species that is consistent with the URF being a protein-coding sequence is that the divergence at silent sites is much higher than the divergence at replacement sites. Table 1 lists the divergence at silent and replacement sites for each pairwise comparison of the 312-bp consensus sequence between *M. domesticus*, *M. caroli*, and *M. platythrix*. There are four times more potential replacement than silent substitutions in these sequences. If nucleotide substitutions were occurring at random, we would expect to find four times more substitutions at replacement sites than at silent sites. Instead, there are roughly equal numbers of substitutions occurring at silent and replacement sites. After correction for multiple hits (37), the divergence at silent sites is about four times greater than the divergence at replacement sites. This suggests that the URF in this repeat family is evolving under selective constraints with respect to mutations causing amino acid replacements. This is a characteristic property of DNA sequences that encode proteins (37, 38).

Recently, DNA sequences from portions of the monkey *L1* (*Kpn* I) family have been published (22, 39). The available sequence includes a region that is homologous to the sequence from *L1Md*-4 and other published *L1Md* sequences (see ref. 19 and references therein). Due to insertions and deletions, this particular representative of the monkey repeat does not have the entire reading frame open as found in the mouse repeat. However, the same 312-bp region from the monkey sequence that was analyzed for the different species of *Mus* can be aligned with the mouse (BALB/c) sequence by the introduction of a 3-bp gap in the mouse sequence. A comparison of the ratio of silent and replacement substitutions between the mouse and the monkey sequences in this region indicates that the homologous sequence in the *L1* family of primates has been evolving under selective constraints similar to the mouse URF. Although this particular representative of the monkey family does not have an open URF due to various frameshifts and terminators, it is not yet known what fraction of the monkey repeats presently carry this sequence as an open reading frame.

Three lines of evidence suggest that the URF has undergone most, if not all, of its evolution under constraints imposed by selection for protein-coding function: (*i*) the presence of the large URF in *L1Md*-4 from *M. domesticus*, (*ii*) the preservation of (a subset of) the same open reading

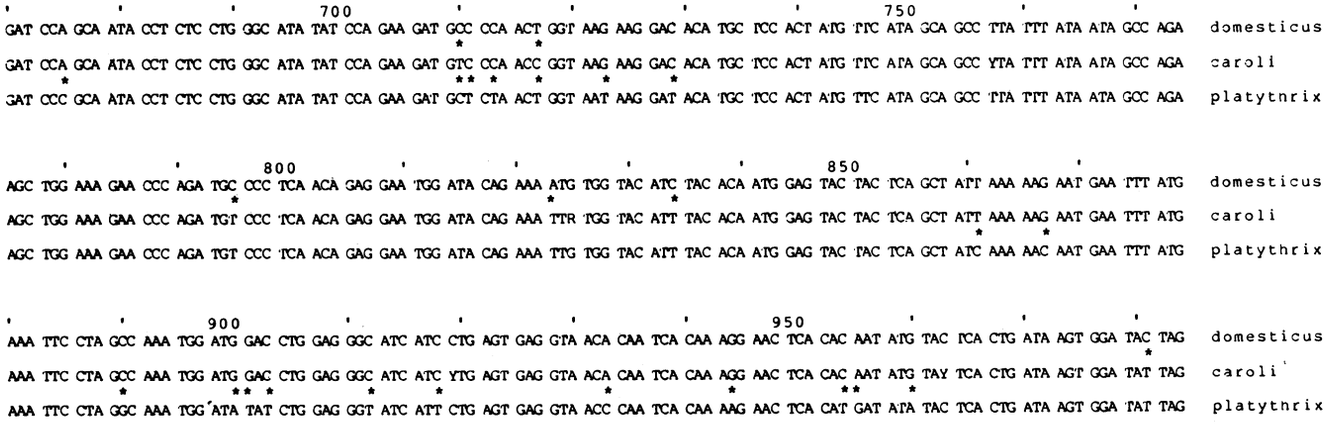

FIG. 4. Consensus sequence of *L1* from *M. domesticus*, *M. caroli*, and *M. platythrix*. The nucleotide found in the majority of sequences within each species is shown for each position. At several positions in the 10 sequences from *M. caroli*, two nucleotides were found in exactly half of the sequences determined. Such positions are indicated by a "Y" (C or T) or an "R" (A or G). In each case, the codon including these ambiguous nucleotides would remain unchanged.

Table 1. Divergence at silent and replacement sites in the URF of the long interspersed repeat

Pair of sequences	Silent sites				Replacement sites				Ratio (S/R)	Corrected ratio (S/R)
	Possible sites	Observed substitutions	Divergence, %	Corrected divergence	Possible sites	Observed substitutions	Divergence, %	Corrected divergence		
M. domesticus, *M. caroli*	62.3	6	9.6	9.4	249.7	2	0.8	1.0	12	9.4
M. domesticus, *M. platythrix*	63	12	19.0	16.0	249	10	4.0	4.5	4.8	3.6
M. caroli, *M. platythrix*	63.3	12	19.0	16.0	248.7	10	4.0	4.5	4.8	3.6
Monkey, mouse	61.2	42	68.5	89.9	250.8	56	22.3	34.0	3.1	2.6

The values for "possible sites" were determined by examining all three possible base substitutions at each position in the sequence and determining whether the altered codon represents a silent (S) or replacement (R) substitution. The tabulated values were obtained by averaging the number of changes in each category for the pairs of sequences and dividing by three because there are three substitutions possible at each site. The "observed substitutions" in each category were scored for each pair of sequences as in Brown *et al.* (37). A crude "divergence" value is calculated as the ratio of observed substitutions to possible sites, expressed as a percent. "Corrected divergence" was determined by using the method of Brown *et al.* (37) to correct for multiple hits at a single site. The number of "ATY" codons found in these sequences is small; thus they were not treated as a separate category. Transitions were considered to account for 70% of the mutations. This number is based on the average number of transitions observed among these three sequences; therefore it may be an underestimate of the actual ratio of transitions to transversions.

frame in diverse species of mice, and (*iii*) the low rate of substitutions at replacement sites compared to silent sites in the *Mus* and monkey sequences.

Possible Roles for the URF. There are many copies (10^4) of this URF in mice and many of them potentially encode a protein. What role might this protein serve? Either it is required for the survival of the organism or it is required for the survival of the repeat. The URF may be part of a larger biologically functional unit such as a transposition element or a virus. Most of the copies present may be remnants of a larger structure that has persisted long after the original dispersal of these sequences. A small number of still functional copies could keep the URF open in the rest of the family by genetic exchange processes.

One possible mechanism for such exchange of genetic information is a specific elaboration of the master–slave concept (40). We imagine a mechanism whereby the protein encoded by the URF provides a *cis*-acting function that facilitates genetic exchange. If the protein is a reverse transcriptase, for example, it could bind to its own messenger immediately after translation to initiate cDNA synthesis. The cDNA could then either insert into the genome, as has been proposed for repeats belonging to the *Alu* family and for processed pseudogenes (12, 14, 41), or act as the intermediate in a gene correction process. Another possible function for the protein would involve directing cDNA or mRNA back to the nucleus.

A model for the genetic exchange involving a cDNA intermediate is consistent with the known truncated structure of the repeat family because premature termination is a feature associated with reverse transcription. Individual family members apparently extend random distances in the 5′ direction from a conserved, A-rich 3′ end point (15, 17). This particular model could explain how the repeat family can avoid being taken over by copies that have acquired termination codons or frameshifts, yet are still transcribed, because the exchange requires a functional protein product of the transcript. A large fraction (81%) of the random isolates contain the open reading frame in the 315-bp region sequenced. This would be predicted by a model in which the product of the URF is required in the genetic exchange process. This is basically a "selfish" model, in which the URF product is necessary to insure survival of the repeat.

Perhaps even more intriguing is the possibility that this repeat and its product are an intimate and required part of the cellular regulatory apparatus. The model proposed above does not preclude the possibility that the repeat exerts effects on the expression of nearby genes. For example, if the repeat contains enhancer sequences originally designed to drive the transcription of its URF, these may stimulate host gene expression as well. Alternatively, the URF product itself may exert some direct effect on gene activity within the repeat or in the surrounding region. If the URF product is important to cell function, then selection could maintain a large number of open reading frames independent of a passive mechanism (such as gene conversion) to generate them. If selection is the basis for the large number of open URFs, then the interesting issue is why does the organism need so many copies of this particular sequence?

In summary, we have identified a long open reading frame in a highly repetitive DNA family. The distribution of mutations within the sequences and the conservation of the open frame between diverse species suggests that the URF has been undergoing most, if not all, of its evolution under selection for a protein product. This long open reading frame behaves as if it currently encodes a functional polypeptide in that it is found on polyribosomes (16) and in RNA transcribed by RNA polymerase II (23). Thus, it seems likely that the *L1* repeat family is not merely junk DNA.

We are grateful to Dr. Michael Potter for *M. platythrix* liver and to S. Hardies and S. Nordeen for helpful discussions. S.L.M. is a fellow of the Jane Coffin Childs Memorial Fund for Medical Research. This work was supported by National Institutes of Health Grants GM21313, GM30180, and AI08998 to M.H.E. and C.A.H.

1. Davidson, E. H., Hough, B. R., Amenson, C. S. & Britten, R. J. (1973) *J. Mol. Biol.* **77**, 1–23.
2. Graham, D. E., Neufeld, B. R., Davidson, E. H. & Britten, R. J. (1974) *Cell* **1**, 127–137.
3. Manning, J. E., Schmid, C. W. & Davidson, N. (1975) *Cell* **4**, 141–155.
4. Eden, F. C., Hendrick, J. P. & Gottlieb, S. S. (1978) *Biochemistry* **17**, 5113–5121.
5. Bonner, J., Garrard, W. T., Gottesfeld, J., Holmes, D. S., Sevall, J. S. & Wilkes, M. (1973) *Cold Spring Harbor Symp. Quant. Biol.* **38**, 303–310.
6. Schmid, C. W. & Deininger, P. L. (1975) *Cell* **6**, 345–358.
7. Britten, R. J. & Davidson, E. H. (1969) *Science* **165**, 349–357.
8. Davidson, E. H. & Britten, R. J. (1979) *Science* **204**, 1052–1059.
9. Doolittle, W. F. & Sapienza, C. (1980) *Nature (London)* **284**, 601–603.

10. Orgel, L. E. & Crick, F. H. C. (1980) *Nature (London)* **284,** 604–607.
11. Georgiev, G. P., Kramerov, D. A., Ryskov, A. P., Skryabin, K. G. & Lukanidin, E. M. (1982) *Cold Spring Harbor Symp. Quant. Biol.* **47,** 1109–1121.
12. Sharp, P. A. (1983) *Nature (London)* **301,** 471–472.
13. Singer, M. F. (1982) *Cell* **28,** 433–434.
14. Jagadeeswaran, P., Forget, B. G. & Weissman, S. M. (1981) *Cell* **26,** 141–142.
15. Voliva, C. F., Jahn, C. L., Comer, M. B., Hutchison, C. A., III, & Edgell, M. H. (1983) *Nucleic Acids Res.* **11,** 8847–8859.
16. Soriano, P., Meunier-Rotival, M. & Bernardi, G. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1816–1820.
17. Fanning, T. G. (1983) *Nucleic Acids Res.* **11,** 5073–5091.
18. Burton, F. H., Voliva, C. F., Edgell, M. H. & Hutchison, C. A., III (1983) *DNA* **2,** 82 (abstr.).
19. Singer, M. F., Thayer, R. E., Grimaldi, G., Lerman, M. I. & Fanning, T. G. (1983) *Nucleic Acids Res.* **11,** 5739–5745.
20. Fanning, T. G. (1982) *Nucleic Acids Res.* **10,** 5003–5013.
21. Kole, L. B., Haynes, S. R. & Jelinek, W. R. (1983) *J. Mol. Biol.* **165,** 257–286.
22. Lerman, M. I., Thayer, R. E. & Singer, M. F. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3966–3970.
23. Shafit-Zagardo, B., Brown, F. L., Zavodny, P. J. & Maio, J. J. (1983) *Nature (London)* **304,** 277–280.
24. Jelinek, W. & Schmid, C. W. (1982) *Annu. Rev. Biochem.* **51,** 813–844.
25. Kan, Y. W. & Dozy, A. M. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 5631–5635.
26. Jahn, C. L., Hutchison, C. A., III, Phillips, S. J., Weaver, S., Haigwood, N. L., Voliva, C. F. & Edgell, M. H. (1980) *Cell* **21,** 159–168.
27. Maxam, A. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
28. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143,** 161–178.
29. Cheng, S.-M. & Schildkraut, C. L. (1980) *Nucleic Acids Res.* **8,** 4075–4090.
30. Heller, R. & Arnheim, N. (1980) *Nucleic Acids Res.* **8,** 5031–5042.
31. Brown, S. D. M. & Dover, G. (1981) *J. Mol. Biol.* **150,** 441–466.
32. Meunier-Rotival, M., Soriano, P., Cuny, G., Strauss, F. & Bernardi, G. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 355–359.
33. Gebhard, W., Meitinger, T., Hochtl, J. & Zachau, H. G. (1982) *J. Mol. Biol.* **157,** 453–471.
34. Wilson, R. & Storb, U. (1983) *Nucleic Acids Res.* **11,** 1803–1817.
35. Dayhoff, M. O., Hunt, L. T., Barker, W. C., Orcutt, B. C., Yeh, L. S., Chen, H. R., George, D. G., Blomquist, M. C. & Johnson, G. C. (1983) *Atlas of Protein Sequence and Structure*, Version 7 (National Biomedical Reserch Foundation, Washington, DC).
36. Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50,** 349–383.
37. Brown, W. M., Prager, E. M., Wang, A. & Wilson, A. C. (1982) *J. Mol. Evol.* **18,** 225–239.
38. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) *Cell* **20,** 555–566.
39. Thayer, R. E. & Singer, M. F. (1983) *Mol. Cell. Biol.* **3,** 967–973.
40. Callan, H. G. (1967) *J. Cell Sci.* **2,** 1–7.
41. Van Arsdell, S. W., Denison, R. A., Bernstein, L. B., Weiner, A. M., Manser, T. & Gesteland, R. F. (1981) *Cell* **26,** 11–17.