Revised genomic consensus for the hypermethylated CpG island region of the human L1 transposon and integration sites of full length L1 elements from recombinant clones made using methylation-tolerant host strains

P.J.Crowther, J.P.Doherty, M.E.Linsenmeyer, M.R.Williamson and D.M.Woodcock*
Molecular Genetics, Peter MacCallum Cancer Institute, 481 Little Lonsdale Street, Melbourne, Victoria 3000, Australia

Received January 24, 1991; Revised and Accepted April 16, 1991 EMBL accession no. X58075

ABSTRACT

Efficient recovery of clones from the 5′ end of the human L1 dispersed repetitive elements necessitates the use of deletion *mcr*⁻ host strains since this region contains a CpG island which is hypermethylated *in vivo*. Clones recovered with conventional *mcr*⁺ hosts seem to have been derived preferentially from L1 members which have accumulated mutations that have removed sites of methylation. We present a revised consensus from the 5′ presumptive control region of these elements. This revised consensus contains a consensus RNA polymerase III promoter which would permit the synthesis of transcripts from the 5′ end of full length L1 elements. Such potential transcripts are likely to exhibit a high degree of secondary structure. In addition, we have determined the flanking sequences for 6 full length L1 elements. The majority of full length L1 clones show no convincing evidence for target site duplication in the insertion site as commonly observed with truncated L1 elements. These data would be consistent with two mechanisms of integration of transposing L1 elements with different mechanisms predominating for full length and truncated elements.

INTRODUCTION

The L1 or LINE-1 family of long dispersed repetitive elements of mammalian genomes has been shown to be related evolutionarily to known retrotransposons in lower eukaryotes (reviewed in ref.1). A general feature of mammalian L1 elements includes an internal region which shows a high degree of sequence homology between species. This region contains 2 putative open reading frames (ORFs), the second of which is thought to encode a reverse transcriptase (2,3). Both the 5′ and 3′ ends of the L1 elements are unique to each mammalian species (1). The 5′ ends of mammalian L1 elements commonly have a high C+G content with an elevated frequency of CpG dinucleotides and, as such, fall under the definition of CpG islands (1,4,5). Housekeeping genes frequently contain CpG islands as part of their control regions where they are typically present in an unmethylated state (4,5). It has been proposed that the CpG island at the 5′ end of the L1 elements is a prime regulatory region for the element and that further colonization of the genome by an L1 element family can occur following the acquisition of a new 5′ control region. In L1 elements of both rat and human genomes, it has been shown that this region is heavily methylated *in vivo*, making them atypical for CpG islands (6,7,8).

The host strains which lack the 5-methylcytosine (5mC) specific restriction systems (*mcrA* and *mcrBC*) need to be used for the efficient and apparently random recovery of clones from the more heavily methylated portions of mammalian genomes. Clones containing this hypermethylated region of the human L1 repeat are typically underrepresented by 30 to 60 fold amongst recombinants recovered using conventional *mcr*⁺ hosts (8,9,10). Further, the clones that are recovered using conventional *mcr*⁺ *E. coli* host strains appear to be biased towards the more degenerate members of the human L1 family (L1Hs) (9) presumably because of this heavy methylation. Such degenerate members of the L1 family typically contain numerous mutations probably accumulated though the spontaneous deamination of methylated cytosines to thymine (9). Here we have used optimally methylation-tolerant host strains to recover a large number of independent clones containing this hypermethylated 5′ presumptive control region of the human L1 elements. We have compared sequence data from these clones to establish a consensus for the 5′ CpG island region free from bias introduced by methylation-restricting host strains. This revised consensus extends into the region of the human L1 element lacking the high frequency of CpG dinucleotides present in the far 5′ end and into the start of the evolutionarily conserved ORF-1 region (11) where our consensus agrees closely with that of the previously published genomic consensus compiled from clones made using conventional *mcr*⁺ hosts. In addition, we have determined the

* To whom correspondence should be addressed

flanking sequences for six full length L1 elements.These predominantly appear to have different characteristics from the sites of insertion of truncated L1 elements.

METHODS AND MATERIALS

Oligonucleotides used in this study were synthesized on a Applied Biosystems DNA Synthesizer using trityl-on synthesis and purified using NEN-SORB columns (NEN-Dupont).

Recovery of clones from the 1.8 kilobase (kb) *Kpn*I region of genomic L1 elements from human spleen DNA has been described previously (8,10). For further sequencing of these clones, the inserts were subcloned into M13 for single-stranded sequencing by standard protocols. The human genomic library in λ EMBL3 was constructed and screened for clones containing the 5′ end of L1Hs as described previously (10). For direct sequencing of λ clones, phage particles were isolated from 400 ml liquid lysates using an initial CsCl step gradient followed by an equilibrium gradient. The DNA was then released from the phage particles by extraction with phenol/chloroform/octanol (25:24:1) (12). Before sequencing, the λ DNA was digested to completion with *Sal*I to separate the inserts from the λ arms and ethanol precipitated. The DNA was dissolved in 10 μl of water containing 100 ng of primer. The tube was boiled for 4 minutes to denature the DNA and immediately snap frozen in a dry ice/ethanol bath. The contents were left to thaw on ice. Immediately upon thawing, the labelling reaction was commenced by the addition of 2 μl 0.1M dithiothreitol, 2 μl Sequenase buffer (5×), 0.5 μl of [^{32}P]dATP, 0.7 μl labelling buffer (1:3 dilution) and 2.5 μl Sequenase enzyme (U.S.B.) (1:8 dilution). The tube was left at room temperature for 5 mins. The contents were then distributed into the 4 tubes containing 2 μl of each termination mix in 4 μl aliquots. These tubes were incubated at 37°C for 5 minutes and then 4 μl of stop solution added. Sequencing gels were performed by standard protocols.

Calculations for optimal folding of potential RNA transcripts was performed by the method of Zucker and Steigler (13) in which a minimum energy of folding is calculated for the whole molecule using the Wisconsin GCG Version 6 of the programme.

RESULTS

Consensus sequence for the 5′ end of L1Hs

From a comparison sequence data for the 5′ and 3′ ends of 18 separate clones from the 1.8 kb *Kpn*I region of the human L1 repeat (L1Hs) recovered using methylation-tolerant host strains, we have previously shown that there are 2 major subfamilies of L1 elements in the human genome as indicated by characteristic bases in certain, apparently progenitor-related positions (9). This 1.8 kb *Kpn*I region starts 100 base pairs (bp) in from the 5′ end of full length L1 elements (14,15). In order to extend the consensus sequence for clones made using *mcr*$^-$ host strains over the whole of the 5′ hypermethylated presumptive promoter region, inserts from 4 of these clones from the major subfamily whose sequence was closest to the consensus were re-cloned into M13 for more efficient single-stranded sequencing. Successive oligonucleotide primers were made as the sequencing progressed to extend the existing sequence. To derive a consensus for the first 100 bases of the L1Hs from the 5′ end to the first consensus *Kpn*I site, a human genomic library constructed in λ EMBL3 using optimally methylation-tolerant host strains was screened for recombinants containing the 5′ end of the 1.8kb *Kpn*I fragment of L1Hs using appropriate oligonucleotide probes (8,10). The DNAs of nine λ phage recombinants containing the 5′ region of L1Hs were sequenced directly without subcloning (Figure 1) and this consensus was combined with the consensus from the M13 clones commencing at the conserved 5′ *Kpn*I site (Figure 2). The comparison of the λ clones also established the genomic consensus for the 5′ end of L1Hs. This coincided exactly with that determined by primer extension assays of full length transcripts from active L1 elements in Ntera-D2 cells (14). Our clones showed a higher level of variability in the 6 to 7 bp at the 5′ end than in the remainder of the first 100 bp (Figure 1). Except for a few presumably progenitor-related sites, there were

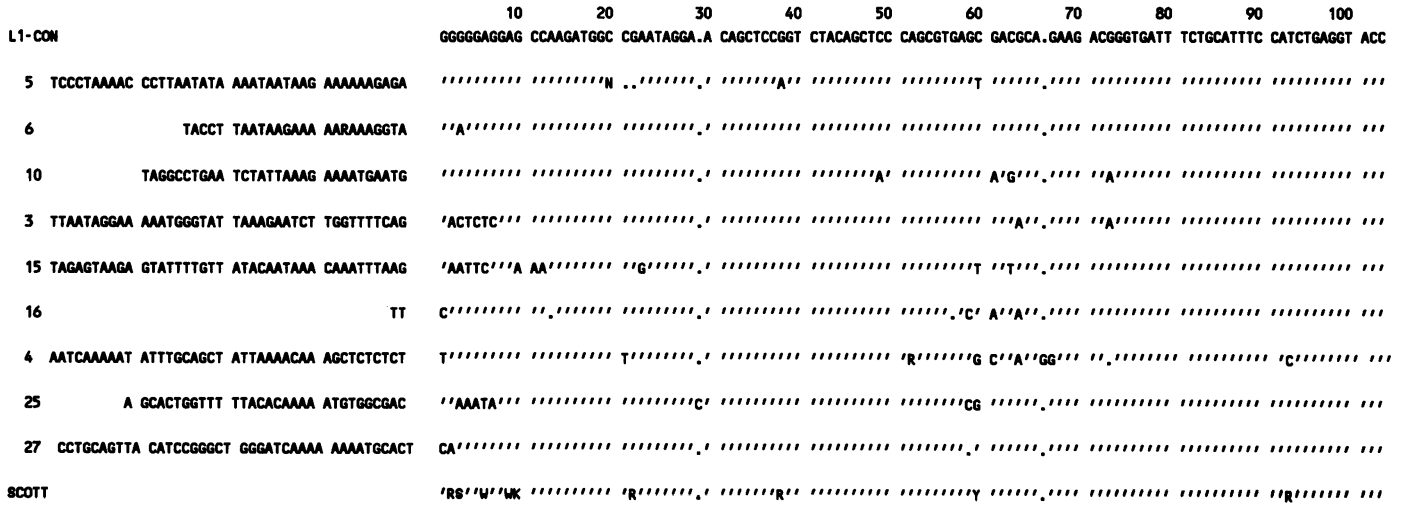

Figure 1. Sequences of the first 100 bases of the 5′ end of L1Hs elements from nine λ clones. The sequence 5′ to base 1 of each clone is also shown to confirm the position of the 5′ end of the genomic consensus. The consensus from the 9 clones is shown on the first line. This consensus is compared with the largest tabulation of L1Hs sequences to date by Scott *et al* (15) (bottom line). The numbers to the left of the figure are numbers for the individual clones. The presence of a (') in each clone indicates concordance with the consensus. Abbreviations for assignment of ambiguous sites are: R=A/G; M=A/C; S=C/G; W=A/T; Y=T/C; K=G/T, P=A/T.

no significant differences beyond the CpG island region between our consensus and the previously published genomic consensus compiled using clones recovered with conventional *mcr*$^+$ host strains (15) (Figure 2). This is presumably due to a significantly lower level of *mcr* restriction of recombinant clones lacking the hypermethylated 5′ end.

Frequency of CpG Dinucleotides

The definition of a 'CpG island' is the frequency of C+G to be greater than 60% and the observed/expected frequency of the CpG dinucleotide of 0.6 (4,5). The consensus for the presumptive control region at the 5′ end of L1Hs derived using methylation tolerant host strains (*mcr*$^-$) conforms to the above definition for the first 500 bases. Figure 3 illustrates the frequency in this region of CpG nucleotides in this consensus and in the previously compiled consensus from *mcr*$^+$ clones (15). The frequency of CpG dinucleotides was significantly higher with the consensus made from the *mcr*$^-$ recombinants (Figure 3). This is also true, but to a lesser extent, for C+G content (not illustrated).

Insertion sites for Full Length L1Hs Elements

A total of 20 λ clones were screened by Southern blotting with four probes covering the entire L1Hs sequence with the exception of the 5′ region of some 100bp (16). Recombinants were considered to contain presumptive full length L1 clones if all of the conserved restriction sites were present and if probes for each of the 4 segments hybridized with fragments of the correct size. Of the recombinants analyzed, only a minority appeared to contain unrearranged full length clones. Eleven clones contained internal deletions or lacked one or more of the 3′ end fragments. Appropriate oligonucleotides were used to obtain the sequences of the ends of the presumptive full length L1 elements together

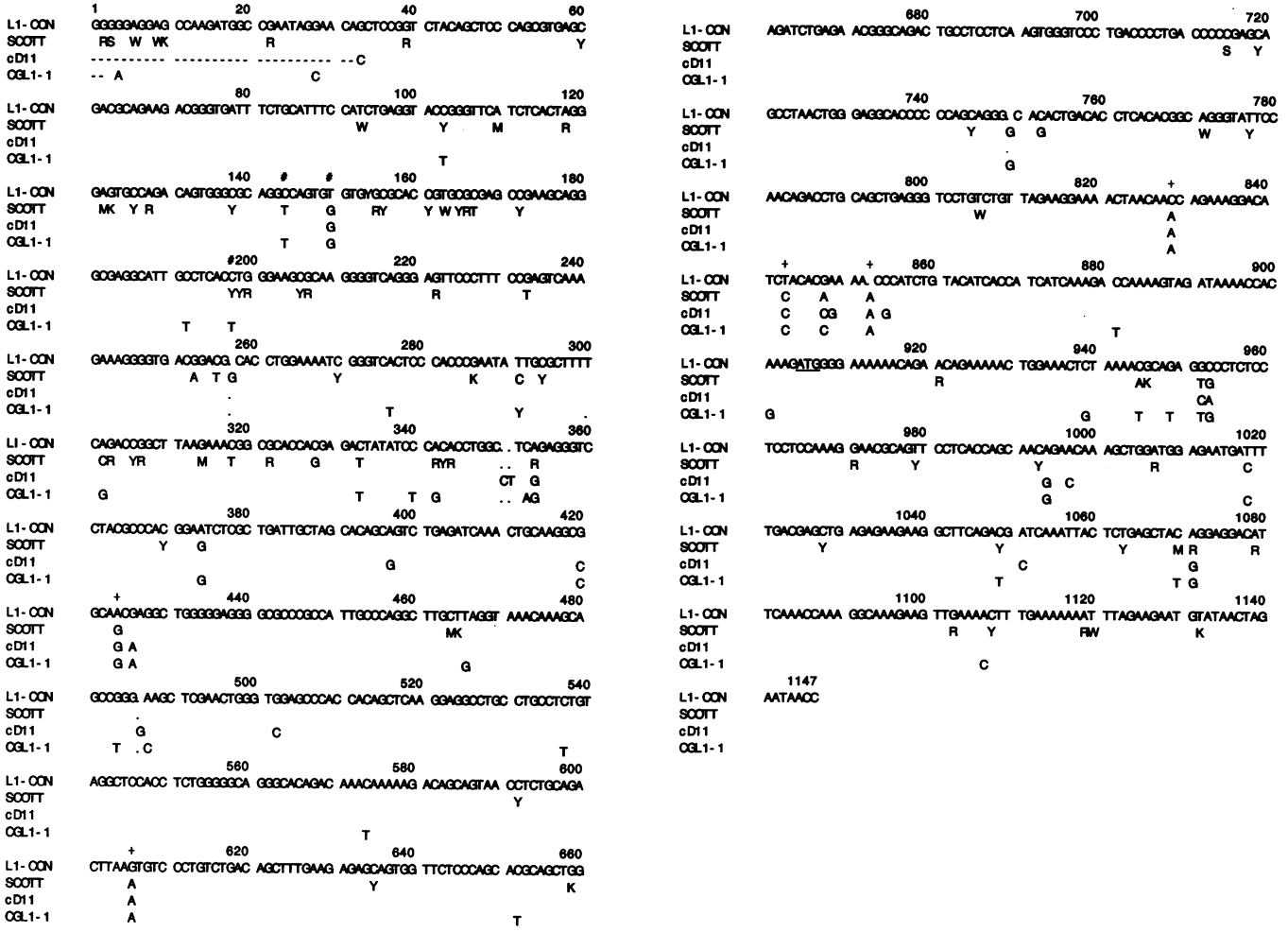

Figure 2. Consensus sequence of the 5′ end of the L1Hs derived from clones in methylation tolerant host strains. The first 100 bases (i.e. up to the *Kpn*I site) is the consensus derived from the nine λ EMBL 3 clones (Figure 1). After the *Kpn*I site the consensus was derived from four clones derived from the most numerous family of genomic L1 elements in the human genome (9). L1-CON represents our 5′ L1 consensus made with *mcr*$^-$ host strains, SCOTT is from the L1Hs consensus of Scott *et al* (15), cD11 is the sequence of this region from an almost full length cDNA-derived L1Hs clone (14), and CGL1−1 is the sequence of a clone from a highly methylated genomic L1 element (17). The Scott L1Hs consensus contains a 132bp insert after base 790 which is not present in the other sequences. The presence of '.' indicates that a base is absent from that sequence at that position while '−' indicates bases missing from the 5′ end of clones. Abbreviations for ambiguities in DNA sequences are as the Figure 1 legend. An '#' above a base indicates sites identifying different L1 families as previously described (9). Other probable family sites are identified by '+'. In our 5′ consensus, the first ATG initiation codon at the beginning of ORF-1 is shown underlined. In the 'ORF-0' region, the non-AUG potential initiation codon, CTG, is at position 82−84 and the TAG stop codon is at position 388−390. A consensus donor splice site is present at bases 247−248 while a possible acceptor site is present inside ORF-1 at bases 269−270.

with that of their flanking DNAs (Figure 4). Some additional apparently full length clones were found by DNA sequencing to be missing the extreme 5′ or 3′ termini and are not included in this comparison. While assignment of the 5′ ends of each L1 element was reasonably unambiguous (Figure 1), determination of the 3′ ends of the L1 elements was more problematic. All the full length clones contained a consensus (or almost consensus) polyadenylation signal sequence (AATAAA) on their 3′ ends. This was followed by 2 to 12 A-residues. We have followed Scott *et al* in defining the 3′ end as the polyadenylation signal (15). In λ clones #10 and #36, the L1 element were flanked by 14 bp and 13 bp repeats respectively. The remaining clones lacked such convincing duplications. The other clones had no sequences immediately adjacent to the 5′ terminus which were repeated on the 3′ side of the element. In λ clone #3, a 6bp sequence (GTTTT) present 2bp from the 5′ end of the L1 sequence was repeated 26 bases downstream from the 3′ end. This was separated from the polyadenylation signal by a mixed A-T sequence which might represent a highly degenerate poly-A tail. The closest sequence feature suggestive of target site duplications in the remaining clones consisted of A-tracts 6 to 8 bases long which were 6 to 9 bases from the 5′ end of the elements with corresponding short oligo-A tracts after the 3′ polyadenylation signal. However, such features could equally well be ascribed being due to insertion in an A-T rich segment without target site duplication.

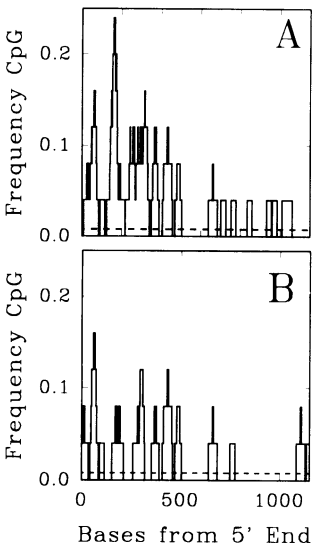

Figure 3. Frequency of CpG dinucleotides in (A) the consensus compiled from clones made using *mcr*$^-$ hosts and (B) the consensus compiled using conventional *mcr*$^+$ hosts (15). The dashed lines indicate the average for the human genome of CpG dinucleotides (35). Frequencies are shown as running averages over 25bp.

DISCUSSION

Comparisons with other Consensuses

The 5′ consensus presented here derived using the methylation-tolerant host strains (deletion *mcr*$^-$) is closest to that of a cDNA-derived clone of a human L1 from N-tera2D1 cells (15) in which there were only 23 changes compared to 1,146 bases in our 5′ L1 consensus (2.1% divergence) (Figure 2). Note that the cD11 clone lacks 32 bp from the 5′ end of L1Hs. Comparison with the largest compilation of L1Hs sequences (from clones made before the introduction of *mcr*$^-$ host strains as improved cloning hosts) (15) is more problematic. Different members of the human L1 element family contain a number of sites of sequence variation in this region which are probably progenitor-related (9). These are sites at which one of two alternate bases are each present in about 50% of clones and in which individual bases correlate with the base assignments at other sites in the individual L1 clones (9) (Figure 2). Beyond the first consensus *Kpn*I site, our consensus was compiled from clones selected from what appears to be the most numerous L1 family. However, the earlier consensus was compiled from the sequences of all available clones without consideration of any potential family groupings. Thus there are a number of sites which are ambiguous in the

Figure 4. Insertion sites of full length L1Hs elements in genomic clones in λ EMBL3. The upper line for each clone shows the DNA sequences 5′ (left) and 3′ (right) of the integrated copy of the L1 element. The lower line shows the sequences at the ends of the integrated L1 element. '{L1}' indicates the major body of the element. The lines indicate the sites of insertion. Flanking repeated sequences are underlined.

earlier (*mcr*$^+$) consensus for this reason. Also, due to the ambiguity of base assignment at other sites in the earlier consensus which in many instances would include the base assignment we were able to make unambiguously using our *mcr*$^-$ clones, it is difficult to calculate a simple numerical value for the divergence between these two consensuses. Possibly a better indication of the significance of the revisions to the consensus for this region can be derived from other types of comparisons. For instance, the *mcr*$^+$ consensus was thought to contain several stop codons in the 5′ region (15) which are absent from the consensus made using the *mcr*$^-$ hosts (see below). Also many CpG dinucleotides are absent in the clones made using *mcr*$^+$ host strains (Figure 3). In addition, the various *mcr*$^+$ genomic clones typically contain short deletions which are not observed in *mcr*$^-$ clones. This latter difference might represent a greater level of sequence degeneracy in the clones recovered using the methylation-restricting strains or might possibly represent cloning artefacts at sites of repair of *mcr* nuclease-induced breaks. Hohjoh *et al* have recently obtained a human genomic L1 clone which is closer in sequence to our consensus in some regions than the consensus derived from clones made using the conventional *mcr*$^+$ hosts (17). These workers postulated that a high CpG dinucleotide content was a characteristic of functional L1Hs elements. In the highly methylated CpG island region of L1Hs, the sequence of the human L1 clone of Hohjoh *et al* is very similar to our consensus for that region (Figure 2). However, overall, it showed a total of 39 changes (3.4%) from our consensus in the whole of this 5′ region with the majority of changes concentrated surprisingly after the 5′ CpG island region in sequences where the *mcr*$^-$ consensus, the *mcr*$^+$ consensus, and the cDNA sequence are all in close agreement. (The concordance in this region between the former 2 consensuses presumably reflects a much lower level of *in vivo* methylation 3′ to the CpG island region.)

RNA pol III	**A block**	**B block**
consensus	5′-{7-16bp}-TGGCNNAGTNGG-{17-60 bp}-GGTTCGANNCC	
		(37)
	(8)	-{ 17 bp }-cGgTCtACAgC
L1Hs consensus	5′-{7bp}-gaGCCAAGaTGG	(56)
		-{ 36 bp }-tGagCGACGCa
	(17)	(56)
" "	5′-{16bp}-TGGCCGAaTAGG-{ 27 bp }-tGagCGACGCa	

Figure 5. Sites of homology to the RNA polymerase III promoter in the *mcr*$^-$ consensus sequence. The RNA polymerase III promoter consensus (20) is shown on the top line. Bases in bold are those which tend to be invariant in the promoter sequence. For the L1Hs consensus, the bases in upper case are present in the consensus, and the bases in lower case are not present in the consensus. The numbers in brackets indicate the position where the homology commences in the consensus in Figure 2. The first potential termination site for RNA polymerase III (an oligo-T tract) is present at base 300 (Figure 2).

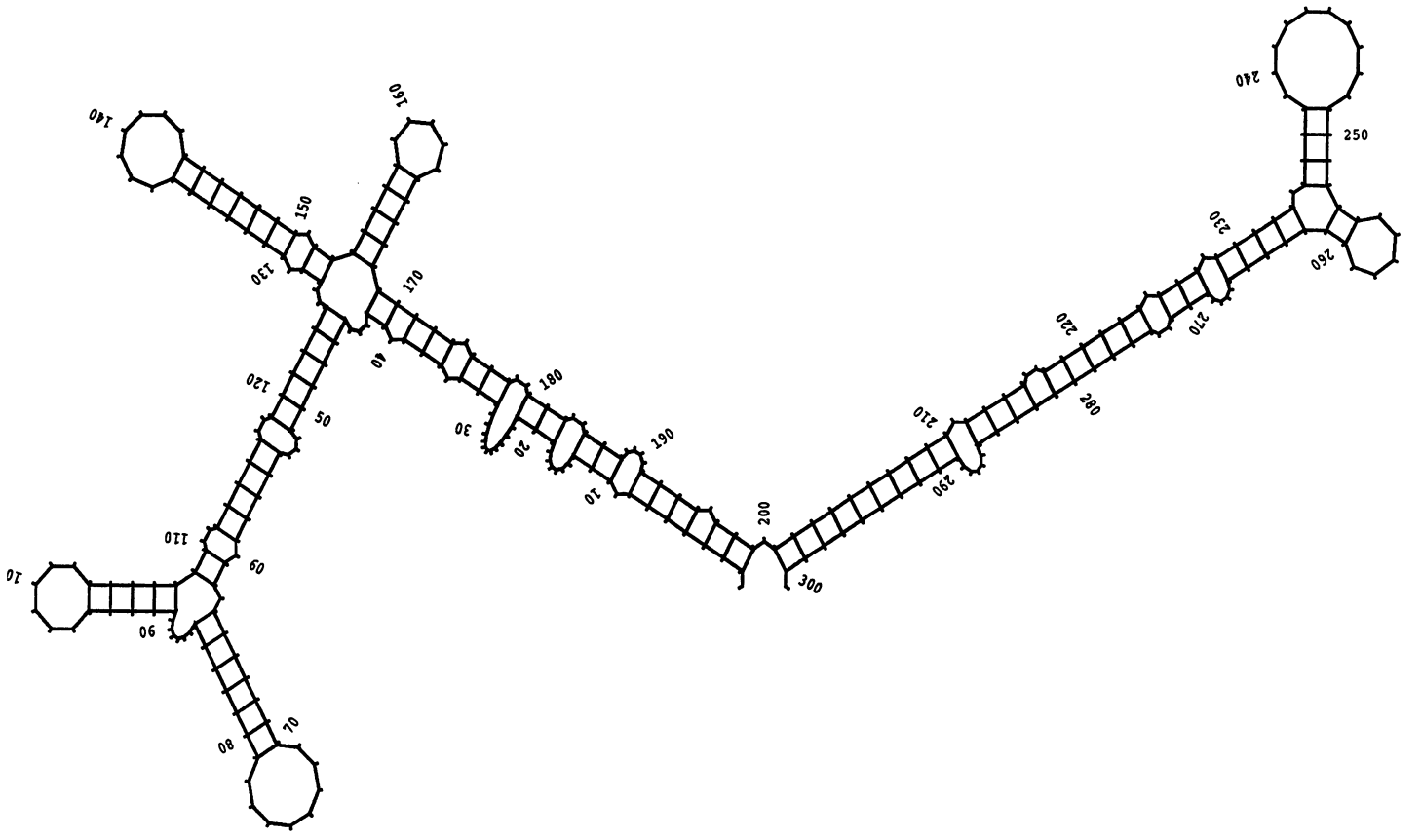

Figure 6. Minimum energy secondary RNA structure calculated by the method of Zucker and Steigler (13) for a potential RNA polymerase III transcript from the 5′ end of L1Hs. The structure for 301 bases from the 5′ end to the first oligo-T tract has a calculated energy of −98.8 kcal.

Features of the *mcr*$^-$ Consensus of the 5′ Region

RNA Polymerase III Promoter

Transcription of L1 elements is thought to proceed via an RNA intermediate as suggested by the evolutionarily conserved homology in ORF-2 to known reverse transcriptases (2,3). For the formation of full length transposition intermediates, human L1 elements must possess an internal promoter otherwise the 5′ end would be lost in a single cycle of transposition (18). A promoter which starts transcription 5′ to the promoter is the split internal promoter of RNA polymerase III. Short interspersed repeated DNA elements (SINES) from different mammalian species contain an RNA polymerase III promoter (19). The consensus sequence for the RNA polymerase III internal promoter has been shown to consist of two blocks (designated A and B) of conserved residues separated by a 17 to 60 nucleotides (20). The 5′ A block is typically 7 to 16 bp from the transcription start site (20). At the 5′ end of our L1Hs consensus, there are two sequence blocks 7bp and 16bp from the 5′ end with 75% and 91% homology respectively with the A block of RNA polymerase III promoter. Homology to the B block of 63% can be found at appropriate distances 3′ to each of these A blocks (Figure 5). Within known RNA polymerase III promoters, there are a number of sites in which a certain base is present in nearly 100% of instances (20). In the A block sequences all these bases are present while, in the B block, 4 out of the 5 bases conform at these conserved positions. Deletion analysis of the 5′ end of the L1Hs revealed that the first 100 bases were critical for transcription (21). This may be due to the presence of a RNA polymerase III promoter in this region.

Potential Secondary Structure of RNA Polymerase III Transcripts

An analysis of the secondary structure of potential RNA polymerase III transcripts from the 5′ end of L1Hs showed a high inherent ability to form stem-loop structures and stacking sites. The potential structure of an RNA transcript from this region is shown in Figure 6. The structure shown is that with the lowest energy computed over the entire molecule. The high degree of secondary structure calculated for this potential transcript is reminiscent of that calculated for a region of the 16S ribosomal RNA (13). The significance of these structures and their actual existence in reality within a cell are purely hypothetical especially since, like tRNA precursors, any such RNA polymerase III transcript might be subject to splicing *in vivo*.

Short Potential ORF in the CpG Island

The revised consensus derived using methylation-tolerant host strains extends into the evolutionarily conserved region of mammalian L1 elements. In agreement with other sequence analyses, the *mcr*$^-$ consensus for this region contains an open reading frames with an ATG initiation methionine codon at base 905 which correlates with the beginning of ORF-1 identified previously (11,15). While the *mcr*$^+$ consensus contains numerous stop codons in the 5′ CpG island region (15), the majority of these stop codons are absent from recombinants recovered from the *mcr*$^-$ hosts. In this 5′ region, the *mcr*$^-$ consensus contains a short (306bp) potential ORF. The 5′ end of this region, however, does not contain an ATG translation initiation codon. However a number of genes which function in mammalian cells have been shown to initiate at codons other than ATG (22). An examples is the c-*myc* oncogene which initiates from a CTG (23). Such a codon is present at base 82 with the next in-frame termination codon at base 388. This raises the possibility that a previously unidentified short reading frame exists at the 5′ end of the L1Hs ('ORF-0'). Alternatively, this short ORF might represent the first exon of ORF-1. Consistent with this hypothesis, the 'ORF-0' region contains a 5′ consensus donor splice site while there is a 3′ acceptor site near the 5′ end of the ORF-1 region (Figure 2). Indeed there might be a functional advantage in limiting transpositional activity through a suboptimal initiation codon and possibly also through additional control of splicing of transcripts such as has been demonstrated in the *Drosophila* P-element transposition system (24,25).

Mechanisms of Transposition

In the absence of experimental systems in which human L1 elements are actively transposing, the molecular mechanisms of L1 transposition are necessarily speculative. The RNA polymerase responsible for transposition of L1Hs has not been identified although, for a long transcript with protein coding capacity, RNA polymerase II would be expected to be involved. However, RNA polymerase II transcripts initiate 3′ to the promoter. Hence, if an RNA transposition intermediate were formed by RNA polymerase II, the transposed copies should all lack the 5′ terminus. However, L1Hs transcripts have been detected by primer extension assay which initiated from the 5′ end of the genomic consensus (14). While the RNA polymerase III promoter present in our consensus could allow transcription from the 5′ end, there are numerous oligo-T termination signals in L1 elements, the first only some 300 bases from the 5′ end. Hence, whatever the potential function of the RNA polymerase III promoter, it does not seem a good candidate for generating full length RNA transposition intermediates. However the *Drosophila* jockey element which is evolutionarily related to mammalian L1 elements appears to be transcribed from an internal promoter by RNA polymerase II (26). This is the only known case of an RNA polymerase II promoter being internal of the sequence to be transcribed. Whether human RNA polymerase II can function similarly with L1Hs remains to be demonstrated. Note that mouse L1 elements differ significantly in this respect from human (and rat) L1 elements in that they contain multiple copies of repetitive sequences containing RNA polymerase II promoter sequences on their 5′ ends which would allow a number of cycles of transposition before all these promoter sequences were deleted (18).

The product of a transcript from an RNA polymerase III promoter at the 5′ end of the human L1 is possibly more likely to have some other function such as a primer for the synthesis of the cDNA strand in retrotransposition. Retroviruses use tRNA molecules to prime the first strand of cDNA synthesis and it has been proposed that the L1 cDNA synthesis is primed by small cellular RNA molecules in a similar fashion (1). Transcripts from the mouse L1 A- and F-type repeats have the potential to form such tRNA-like structures (Bertling *et al*, quoted in ref.1). The potential to form similar structures have also been identified in transcripts from evolutionarily related transposable elements (*Drosophila I* and trypansome *ingi* elements) (1). The sequence of our 5′ consensus would be consistent with it encoding an RNA polymerase III transcript which has a high potential to form secondary structures (Figure 6). Such an RNA molecule might provide the primer for the synthesis of the first strand of cDNA synthesis by the reverse transcriptase encoded by ORF-2 of the L1Hs through pairing of the oligo-U tail from the RNA

polymerase III termination site with the poly-A tail of the full length L1Hs transcript.

Insertion Sites

The insertion sites of human L1 elements found in this and other studies are in A-T rich tracts (1,27). In this, L1 transposable elements are not in any way unusual. Many studies have shown that integration of any new piece of DNA into some site in a chromosome usually occurs in an A-T rich region. For instance, the integration of transfected DNA into mammalian cells occurs in A-T rich regions apparently at random with little or no sequence homology at the insertion site (28,29,30). Excised copies of Simian Virus 40 in rat cells which has been shown to integrate in A-T rich region (31). Transposons from a variety of species, including Drosophila (*copia*), nematodes (Tc1), rat (L1Rn), mouse (L1Md) and human (*Alu*), all integrate into A-T rich regions of the genome (1). The A-T rich sequences found flanking full length L1 elements are also reminiscent of 'matrix associated regions' (MARs), segments of DNA with high affinity for the nuclear scaffold (32). Indeed, several L1 elements have been shown to be inserted into MARs in the human IgH locus (33). MARs have also been shown to contain high concentrations of topoisomerase II sites (32), suggestive of a potential role for topoisomerase II cleavage in the generation of insertion sites for L1 transposition intermediates.

However, while the A-T rich nature of the integration sites of L1 elements do not appear in any way exceptional, the majority of integration sites reported for L1 elements differ from the integration sites of extrachromosomal DNAs introduced into the cell by transfection in that most L1 elements appear to be flanked by short direct repeats (SDRs) of 12 and 13 bases long (reviewed in ref.1). This is unusual in that integrated copies of exogenous DNAs have typically been inserted without forming target site duplications and with at most short regions of patchy homology with the original sequence of the insertion site (28,29,30). To our knowledge, the only published example of a site of insertion of a full length human L1 element is that for clone TGb41 (34). Taking data for this clone together with our data (Figure 4), it appears that the flanking sequences of the majority of full length L1Hs elements lack SDRs. Only 2 of 7 elements studied had sequence elements (of 14bp and 13bp) immediately flanking the 5′ end which were repeated at the 3′ end after a variable length of A-rich sequence. The other 5 elements did not have repeat units immediately flanking the L1Hs element except for short oligo-A or oligo-T tracts which were not immediately adjacent to the 5′ terminus of the L1 repeats. Such oligo-A and oligo-T tracts are most likely to be present simply due to the A-T rich nature of insertion sites. Thus the predominant form of insertion site for full length transposition intermediates of L1 elements appears to be equivalent to that observed with double stranded transfected DNAs. This is probably not unexpected if L1 transposition intermediates integrate via a non-specific mechanism such as operates on any extrachromosomal double-stranded DNAs. In this regard, the presence of SDRs flanking L1 elements formed from defective (truncated) transposition intermediates seems to the exceptional occurrence.

If any double stranded DNA is normally integrated without SDRs, this would suggest that integrated copies of defective L1 transposition intermediates which give rise to truncated copies with flanking SDRs were not in the form of double-stranded DNA when they were integrated into the chromosome. The most likely alternative structures for such defective transposition intermediates would be incomplete products of the reverse transcriptase such as RNA/DNA duplexes, incompletely formed and only partially double-stranded DNA intermediates, and possibly even single-stranded DNAs.

ACKNOWLEDGEMENTS

We would like to thank Joy Vearing for synthesis of oligonucleotides, Dr. Linda Stern for computer analysis of potential RNA secondary structures, and Dr. Miriam Ford for advice and discussion.

REFERENCES

1. Hutchinson,C.A.,III, Hardies,S.C., Loeb,D.D., Shehee,W.R. and Edgell,M.H. (1989) In Berg,D.E. and Howe,M.M. (ed.) Mobile DNA. American Society for Microbiology, Washington, D.C., pp.593–617.
2. Fawcett,D.H., Lister,C.K., Kellett,E., Finnegan,D.J. (1986) *Cell*, **47**, 1007–1015.
3. Hattori,M., Kuhara,S., Takenaka,O. and Sakaki,Y. (1986) *Nature*, **321**, 625–628.
4. Bird,A.P. (1986) *Nature*, **321**, 209–213.
5. Gardiner-Garden,M. and Frommer,M. (1987) *J. Mol. Biol.*, **196**, 261–282.
6. Furano,A.V., Robb,S.M. and Robb,F.T. (1988) *Nucl. Acids Res.*, **16**, 9215–9230.
7. Nur,I., Pascale,E. and Furano,A.V. (1988) *Nucl. Acids Res.*, **16**, 9233–9251.
8. Woodcock,D.M., Crowther,P.J., Diver,W.P., Graham,M., Bateman,C., Baker,D.J. and Smith,S.S. (1988) *Nucl. Acids Res.*, **16**, 4465–4482.
9. Crowther,P.J., Cartwight,A.L., Hocking,A., Jefferson,S., Ford,M.D. and Woodcock,D.M. (1988) *Nucl. Acids Res.*, **17**, 7229–7239.
10. Doherty,J.P., Graham,M.W., Linsenmeyer,M.E., Crowther,P.J., Williamson,M. and Woodcock,D.M. (1991) *Gene*, **98**, 77–82.
11. Liebold, D.M., Swergold, G.D., Singer, M.F., Thayer, R.E., Dombroski, B.A. and Fanning, T.G. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 6990–6994.
12. Ausubel,F.M., Brent,R., Kingston,R.E., Moore,D.D., Smith,J.A., Seidman,J.G. and Struhl,K. (1989) Current Protocols in Molecular Biology. John Wiley and Sons, New York.
13. Zucker,M. and Steigler,P. (1981) *Nucl. Acids Res.*, **9**, 133–148.
14. Skowronski,J., Fanning, T.G. and Singer,M.F. (1988) M19503 *Mol.Cell Biol.*, **8**, 1385–1397.
15. Scott, A.F., Schmeckpeper,B.J., Abdelrazik,M., Comey,C.T., O'Hara,B., Rossiter,J.P., Cooley,T., Heath,P., Smith,K.D. and Margolet,L. (1987) J03034 *Genomics*, **1**, 113–125.
16. Shafit-Zagardo,B., Brown,F.L., Maio,J.J. and Adams,J.W. (1982) *Gene*, **20** 397–407.
17. Hohjoh,H., Minakami,R. and Sakaki,Y. (1990) X52230 *Nucl. Acids Res.*, **18**, 4099–4104.
18. Loeb,D.D., Padgett,R.W., Hardies,S.C., Shehee,W.R., Comer,M.W., Edgell,M.H. and Hutchinson,C.A.,III (1986) *Mol. Cell. Biol.*, **6**, 168–182.
19. Sakamoto,K. and Okada,N. (1985) *J. Mol. Evol.*, **22**, 134–140.
20. Galli,G., Hofstetter,H. and Birnsteil,M.L. (1981) *Nature*, **294**, 626–631.
21. Swergold,G.D. (1990) *Mol. Cell. Biol.*, in press.
22. Peabody, D.S. (1989) *J. Biol. Chem.*, **264**, 5031–5035.
23. Hann,S.R., King,M.W., Bentley,D.L., Anderson,C.W. and Eisenman,R.N. (1978) *Cell*, **52**, 185–195.
24. Craig,N.L. (1990) *Cell*, **62**, 399–402.
25. Seibel,C.W. and Rio,D.C. (1990) *Science*, **248**, 1200–1208.
26. Mizrokhi,L.J., Georgieva,S.G. and Ilyin,Y.V. (1988) *Cell*, **54**, 685–691.
27. Usdin,K. and Furano,A.V. (1989) *J. Biol. Chem*, **264**, 20736–20743.
28. Murane,J.P., Yezzi,M.J. and Young,B.R. (1990) *Nucl. Acids Res.*, **18**, 2733–2728.
29. Kato,S., Anderson,R.A. and Camerini-Otero,R.D. (1986) *Mol. Cell. Biol.*, **6**, 1787–1795.
30. Kato, S., Anderson, R.A. Cemerini-Otero, R. (1986) *Mol. Cell. Biol.*, **6**, 1787–1795.
31. Bullock,P., Forrester,W. and Botchan,M. (1984) *J. Mol.Biol.*, **174**, 55–84.
32. Sperry, A.O., Blaquez, V.C. and Garrard, W.T. (1989) *Proc. Nalt. Acad. Sci. USA*, **86**, 5497–5501.
33. Cockerill, P.N. (1990) *Nucleic Acids Res.*, **18**, 2643–2648.
34. Fujita,A., Hattori,M., Takenaka,O. and Sakaki,Y. (1987) *Nucl. Acids Res.*, **15**, 4007–4020.
35. Woodcock,D.M., Crowther,P.J. and Diver, W.P. (1988) *Biochem. Biophys. Res. Comm.*, **145**, 888–894.