

Maxine

Please look at this if you
have time today so I can
send it back for corrections.
The whole thing has to be
in final by Feb. 9.

RESEARCH PROGRAM

Maxine Singer, Ph.D.

Beverly
301-402-3095
fax

Sharon:

Please FAX to Beverly. I've made
some minor editing changes. Otherwise,
it's OK

Maxine

Background

The major emphasis of our group is on the LINE-1 (L1Hs) elements in the human genome. The haploid human genome is estimated to contain about 4000 full-length L1Hs elements and approximately 100,000 truncated elements. The full-length elements are 6000 bp; truncated elements range from tens of base pairs to 5000 bp or more. Among the full-length elements are one or more active, transposable elements, and the truncated elements appear to be the result of transpositions accumulated over time. *inactivated*

Active L1Hs elements fall into the class of non-LTR retrotransposons, also referred to as polyA⁺ retrotransposons because the 3' end of the coding strand of L1Hs elements terminates in a stretch of dA residues of variable length. The coding strand of L1Hs elements contains two potential open reading frames of approximately 1 kbp (ORF1) and 4 kbp (ORF2), respectively. These are in the same reading frame but are separated by an 'interorf' region containing two or more in-frame stop codons, depending on the particular element. *?*
Polypurine
stretch

LINE-1 elements are ubiquitous in placental mammals and marsupials and account for as much as 5 percent of these genomes. Related non-LTR retrotransposons are known to occur in a variety of plant and invertebrate genomes. Early work in this and other labs first suggested that LINE-1 elements might represent transposable elements because (1) they are often surrounded by short direct repeats that were either known to be or presumed to be target site duplications, and (2) alleles of unrelated genomic segments could differ by the presence or absence of a LINE-1 element. Active transposition of L1Hs has now been demonstrated. Several examples of mutagenesis in the human genome have been observed and suggest that both germline and somatic cell transposition are contemporary processes. *a*
a
insertional

Using the distinctive sequence variations in a truncated L1Hs element recently transposed into a human Factor VIII gene (and thereby causing hemophilia A), Kazazian and his colleagues identified and cloned a full-length element likely to be the source of the insertion. This L1Hs, L1.2B, is a member of a small class of L1Hs elements that we previously identified as being transcribed in human teratocarcinoma cell lines. In several of the members of this class, both ORF1 and ORF2 are fully open. This is in contrast to most L1Hs elements, in which the ORF's are disrupted by rearrangements, random stop codons, and other mutations. L1.2A, an allele of L1.2B at the LRE-1 locus on chromosome 22q11, also has fully open ORF's and differs from L1.2b by only three base pairs in the second frame. Experiments carried out in Kazazian's laboratory have demonstrated (in yeast) that the second open reading frame in L1.2A encodes an active reverse transcriptase. *B*

Aims

In analogy with the LTR retrotransposons and retroviruses, the current model for the mechanism of L1Hs transpositions involves (1) synthesis of full-length polyadenylated L1Hs RNA, (2) translation of this RNA into proteins, one or more of which is required for

transposition, including the reverse transcriptase, (3) reverse transcription of the RNA by the L1Hs encoded enzyme, and (4) insertion of the cDNA into staggered chromosomal breaks.

Much of our effort is related to investigating various aspects of this model. Earlier we investigated the specific transcription of L1Hs elements. Transcripts of L1Hs are abundant in the nuclei of many human cell types; however, these appear to result from the transcription of L1Hs elements inserted into a variety of unrelated transcription units. Most cell types do not have detectable cytoplasmic full-length polyadenylated L1Hs RNA, the expected specific transcript and messenger RNA. The one exception was in human teratocarcinoma cells in culture. Characterization of these cytoplasmic RNA's by cDNA cloning and other techniques indicated that they had been initiated at residue 1 of the full-length L1Hs elements, represented complete transcripts of the entire element, and were largely transcribed from a particular class of L1Hs elements defined by specific characteristic sequences (as mentioned above). These experiments raised the question of the position of transcriptional regulatory elements that could direct the initiation of transcription at residue 1 and also account for the cell-type specificity of transcription. This question is underscored by the fact that, as far as is known, the DNA sequences surrounding the various transcribed L1Hs elements are likely to be quite different.

We are also studying two aspects of L1Hs translation. First, we are interested in identifying and characterizing the putative proteins encoded by L1Hs within human cells. Second, we are interested in the mechanism of translation. Two aspects of L1Hs structure suggest that translation is unusual: (1) the unusually long (900 base) 5' untranslated region (UTR) and (2) the several in-frame stop codons in the interorf region between ORF1 and ORF2.

Accomplishments

Synthesis and Characterization of p40 in Human Teratocarcinoma Cells

The last few years have seen substantial progress in our knowledge of the protein encoded by ORF1 of L1Hs. Much of this work depends on the polyclonal antibody we prepared to a *trp* E-ORF1 fusion protein synthesized in *E. coli*. Western blotting experiments using this antiserum detect, in human cells, an endogenous protein that migrates in SDS-PAGE with a mobility close to that predicted (40 kDa). The protein, called p40, was detected in extracts of human teratocarcinoma cell lines (NTera2D1 and 2102EP) and in the choriocarcinoma cell line JEG3. In contrast, only very low amounts of protein of the same mobility were detectable by the antiserum in HeLa and 293 cells; the significance of these bands is unknown. No reactive material was detected in HL60 cells. Both *in situ* immunocytochemical analysis and cell fractionation experiments indicate that the bulk and perhaps all of the p40 in teratocarcinoma cells is located in the cytoplasm. Two-dimensional gel electrophoresis (carried out by J.E. Celis) indicates that the size of p40 is 40.3 kDa. The protein is heterogeneous with respect to pI. Some of it has a pI of 10.2 (10.3 predicted) and some has a range of lower pI values. The heterogeneity may be explained, as least in part, by

the fact that p40 is phosphorylated, as indicated by the effect of phosphatase treatment on electrophoretic mobility and the incorporation of ^{32}P from γ -labeled ATP supplied to teratocarcinoma cells.

When NTera2D1 cells were transfected with the plasmid p3LZ, which carries the ORF1 of the previously isolated cD11 cDNA, an additional immunoreactive band was detected with a mobility in SDS-PAGE slightly less than that of the endogenous p40. This experiment confirmed the characterization of the cDNA and the specificity of the antiserum, but also raised the question of why the cDNA produces a p40 of lower mobility than endogenous protein. We then tested a plasmid containing the ORF1 from the cloned genomic L1Hs, L1.2A (from Kazazian and coworkers). *In vitro* translation experiments had indicated that L1.2A produced a p40 with a faster mobility than cD11. The p40 produced from the L1.2A sequence upon transfection into NTera2D1 cells also had a faster mobility than the cD11 and co-electrophoresed with the endogenous p40. This finding was consistent with the conclusion of Kazazian and coworkers that L1.2A is likely to be an active L1Hs retrotransposon. d

The ORF1's in cD11 and L1.2A predict polypeptides of the same length: 338 amino acids. There are 11 amino acid differences predicted. We then investigated whether these differences, or related changes in post-translational modifications, could account for the differing mobilities. To do this, we made constructs in which three different regions of cD11 coding sequences—an amino-terminal portion, the center, and a carboxy-terminal portion—were substituted for the corresponding regions in the L1.2A ORF1. These constructs were transfected into NTera2D1 cells and the mobility of the p40 product investigated by Western blotting. The results indicated that one or more of the four amino acid differences between amino acids 97 and 207 are at least in part responsible for the difference in mobility. NTera
↑ ↑

In vitro translation experiments (reticulocyte lysates and wheat germ) confirmed that the several cDNA clones, representing teratocarcinoma cell full-length cytoplasmic, polyadenylated L1Hs RNA's that have open ORF1's encode p40's with different mobilities in SDS-PAGE. This, and the fact that probably only one encodes a p40 with a mobility like that of endogenous p40 suggests that the bulk of the L1Hs RNA's represented in the 19 cDNA's studied do not serve as functional, productive mRNA's. SDS-
↑ ↑

An interesting feature within the central region of p40 is a potential leucine zipper structure, very similar in important residues to the well-characterized GCN4 leucine zipper. No basic region precedes the zipper segment in p40. Moreover p40 is cytoplasmic and does not appear to be a DNA-binding protein (preliminary experiments). Cross-linking by glutaraldehyde of p40 present in teratocarcinoma cell extracts and of p40 synthesized in *E. coli* indicates that it can form homomultimeric complexes. er

Although ORF1 occupies a position in L1Hs that is analogous to that of gag and gag-like polypeptides in LTR retrotransposons and retroviruses, it has no homology to these proteins. Moreover, as indicated above, p40 does not appear to be subject to proteolytic maturation as

are the primary translation products of gag coding regions. Thus, it is difficult even to speculate on the significance of p40, if any, to the transposition process at this time.

Earlier experiments in this laboratory indicated that the full-length polyadenylated RNA present in the cytoplasm of NTera2D1 cells was undetectable after the cells were induced to differentiate with retinoic acid. Therefore it was surprising to find that p40 was present in approximately the same concentration in undifferentiated cells and in cells at various times after the initiation of retinoic acid treatment. Moreover, pulse-label experiments with ³⁵S-methionine indicate that p40 (isolated by immunoprecipitation) synthesis continues for at least 7 days after the beginning of retinoic acid treatment. Differentiation proceeded as expected, as indicated by several markers. In view of the unexpected nature of these results, we reinvestigated the presence of the full-length polyadenylated L1Hs RNA in NTera2D1 cells before and after retinoic acid treatment, using PCR rather than the Northern blotting technique used earlier. The new experiments revealed that the L1Hs RNA was detectable up to 14 days after retinoic acid treatment was initiated. The PCR experiments do not afford quantitative estimates. We have considered various explanations for what may only be an apparent difference between the new and earlier experiments. Thus, it may be that very low levels of RNA persist, levels that were undetectable by the Northern blotting technique. Furthermore, because there is evidence (see above) that at most only a small number of the RNA chains cloned as cDNA's can represent productive messenger RNA's, the bulk of the RNA detected in undifferentiated cells may not contribute to p40 synthesis.

Transcription of L1Hs Elements

A typical upstream RNA polymerase II regulatory region would be lost during cycles of transposition of non-LTR retrotransposons, in contrast with the situation that holds for LTR retrotransposons. Yet, there is evidence to suggest that already transposed L1Hs elements can be actively transposed and from different genomic locations. Therefore we investigated the possible presence of internal polII regulatory elements within L1Hs elements known to be transcribable. For this purpose, we fused the *lac Z* reporter gene, in frame, after the 15th codon of the ORF1 of cD11 and investigated expression of β -galactosidase in a construct containing a full-length 5'-UTR and various deletions thereof after transfection into NTera2D1 cells. The β -galactosidase activity was assayed both by staining cells with a chromogenic substrate and by enzyme assays on cell extracts. These experiments indicated that the 5'-UTR of the L1Hs is sufficient to permit a substantial level of expression in human teratocarcinoma cells, but not in, for example, HeLa cells, and to specify the transcriptional start site at residue 1 of L1Hs, upstream of the regulatory sequences. Thus, the 5'-UTR has the regulatory elements sufficient for cell-type-specific transcription and for start-site specification. The cell-type specificity of these experiments was consistent with what is already known about the specificity of cells for L1Hs full-length RNA and for the presence of p40.

Deletion analysis revealed that the sequences most critical for transcription are located within the first 100 bp of the L1Hs 5'-UTR. However, sequences spread over the first approximately

668 bp are important for maximal expression rates. We are now studying these regions in more detail.

Deletion of the first 100 bp from the reporter gene construct resulted in a 300-fold reduction in expression of β -galactosidase. Deletion of the first 11 bp has little or no effect, but deletion up to bp 18 reduces reporter gene expression about fivefold, while deletion up to bp 32 has little additional effect. Oligonucleotides containing this region of L1Hs form a specific complex with a nuclear protein extracted from Jurkat and NTera2D1 cells. This complex is ablated by antiserum to the transcription factor called, among other things, YY1. The L1Hs oligonucleotide, as well as an oligonucleotide carrying a related sequence from an unrelated gene known to be responsive to YY1, compete with the L1Hs oligonucleotide for binding to the YY1 in NTera2D1 and Jurkat extracts as well as to YY1 synthesized in *E. coli*. Thus, one of the transcriptional regulatory factors important for L1Hs transcription is likely to be YY1. However, YY1, being ubiquitous, cannot by itself account for the cell-type specificity of L1Hs expression. It is interesting that YY1 is known to be important in the transcriptional regulation of other genes with downstream promoters (*e.g.*, the rpL30 of mice).

A series of small deletions, replacements, and insertions between bp 71 and 102 in the L1Hs 5'-UTR have indicated that this region too is important for transcription. For example, deletion of bp 73 through 80 or insertion of 4 bp after bp 99 reduces expression by about twentyfold. Current ongoing band-shift experiments indicate that the sequence between bp 80 and 85 binds to a protein in NTera2D1 extracts. A sequence at bp 426 through 452 has been shown to bind to Sp1, another well-characterized and ubiquitous transcriptional regulatory factor. DNA footprinting experiments revealed protection of sequences from 504 to 526 by one or more binding proteins. The relatively large region between bp 385 and 525 enhances gene expression (from an SV40 promoter) about tenfold in a manner that is independent of either orientation or location. Altogether, regulation of L1Hs transcription appears to involve a substantial number of regulatory elements spread over a long distance within the element.

Translation of L1Hs elements

Several features of L1Hs suggest special questions regarding translation of ORF1 and ORF2. First, there is the very long 5'-UTR, which is GC-rich. Our computer analysis indicates that the 900-bp segment has the potential to form stable secondary structures. Moreover, each of the L1.2 alleles and the cDNA's characterized thus far have at least one AUG codon in the 5'-UTR upstream of the first methionine codon in ORF1, the codon that our evidence suggests initiates translation of p40. The upstream AUG's could initiate short open reading frames (3 to 20 codons). These structural considerations suggest that translation of ORF1 might be impeded if a scanning 40S ribosome, starting at the 5' end, had to traverse the whole 5'-UTR. Indeed, *in vitro* translation of ORF1 from an mRNA with a very short leader sequence is appreciably more efficient than from L1Hs RNA. Nevertheless, we know from the experiments described above that ORF1 is translated both *in vitro* and in cells, and we are using both these systems, with cloned segments, to investigate the translational process.

With respect to ORF1, an L1Hs element (L1Hs-1.1) with a single-base deletion after the first AUG in ORF1 fails to translate p40, confirming the conclusion that translation starts at this point. When a very stable hairpin structure is introduced into the region of the 5'-UTR, beyond what is important for transcription, ORF1 translation is decreased about fivefold. Similar results were obtained for translation after transfection of plasmids into NTera2D1 cells.

While ORF1 and ORF2 are in the same frame, they are separated by an inter-ORF region of 33 bases bracketed by two conserved in-frame stop codons. We carried out *in vitro* translation experiments using a variety of derivatives of the L1Hs sequence to examine properties of translation. ORF2 polypeptides are synthesized in the same way and relatively efficiently regardless of whether ORF1 is present or being translated. No p40-ORF2 fusion protein was detected in any experiments. The size of the longest polypeptides synthesized was consistent with initiation at the first AUG codon in ORF2. These experiments suggested that ORF2 is translated by reinitiation of translation after cessations at the end of ORF1. Consistent with this conclusion is the fact that the stable stem-loop (see above) that inhibits ORF1 translation had no discernible effect *in vitro* on the translation of ORF2.

The translation of the marker gene β -galactosidase fused into the beginning of ORF2, in frame, was also examined after transfection of plasmids into NTera2D1 cells. The enzyme was not detectable in cell extracts. However, by using the *in situ* staining method, we estimated that ORF2 is translated at least 100 times less efficiently than ORF1. Surprisingly, with cells transfected with the construct that contains the stem-loop in the 5'-UTR (see above), two to five times as many cells showed enzyme activity than in the standard L1Hs construct with reporter gene. Thus, the same stem-loop that decreases translation of ORF1, stimulates translation of ORF2. These results are incompatible with a reinitiation of translation by ribosomes that have already translated ORF1. Rather, they suggest that translation of ORF2 results from internal initiation by attaching ribosomes.

Future Plans

Our general plan is to continue along the lines of the current work described above. The focus will be on the following aspects.

Transcription

Dissection of the complex regulatory region in the L1Hs 5'-UTR will be an important activity. We will aim to understand which of the several regulatory elements in the 5'-UTR are responsible for the cell-type specificity of transcription of L1Hs and also the mechanism by which the transcriptional start site is set at residue 1 by downstream regulatory elements. The study of transcriptional regulatory elements is a major endeavor in current molecular biology. We will therefore aim to keep our work targeted on those aspects that are specific for L1Hs. The major change in approach will be to set up *in vitro* transcription systems in

order to sort out the multiple regulatory elements. We will also continue to characterize the protein that binds to residues 80-85 in the 5'-UTR and to study the apparent enhancer region contained between residues 385 and 525 of the 5'-UTR. We will also devote some attention to learning whether important elements reside in the region between 150 and 385, a segment that has not previously been investigated.

The properties of p40, synthesized in *E. coli*. and in NTera2D1 cells, will be studied. In particular, we would like to understand the nature of the multimeric complexes that are formed, including the possibility that a coiled-coil formed through the leucine zippers may be stabilized by disulfide bonds. Because no proteins in the GenBank have a marked similarity to p40, it is difficult to discern clues as to its function, but by investigating its structural aspects in the cell, we may get some ideas.

A major thrust will be to synthesize polypeptides encoded by ORF2 in *E. coli*. We have already made fusion proteins of some sections of ORF2 and used these to make antibodies. We would like, in particular, to prepare molecules that contain the active reverse transcriptase known to be encoded in ORF2 and to study the properties of this enzyme.