Localization of the L1Md family of repeated sequences in the mouse albumin- α -fetoprotein gene complex

Robin Cooper, Cynthia E. Herzog, Ming-Liang Li, William F. Zapisek¹, Peter R. Hoyt, Harry Ratrie III and John Papaconstantinou*

Department of Human Biological Chemistry and Genetics, Division of Cell Biology/F43, University of Texas Medical Branch, Galveston, TX 77550, and ¹Biochemistry Curriculum, Canisius College, Buffalo, NY 14208, USA

Received 22 May 1984; Revised and Accepted 17 July 1984

ABSTRACT

Studies on the β -globin gene complex in the mouse have demonstrated the existence of repeated DNA sequences interspersed throughout the intergenic regions (1,2). These sequences are members of families of middle repetitive sequences and have been mapped to specific intergenic sites in the 60 kbp β -globin complex. In this study we present evidence that members of this middle repetitive family of DNA sequences, the L1Md family, are interspersed throughout the mouse albumin and α -fetoprotein gene complex. Unlike those of the β -globin complex, all of which are found in the intergenic regions, these sequences are localized within intron 12 of the albumin gene and intron 3 of the AFP gene as well as twice in the 13.5 kbp intergenic region that links the albumin gene to the AFP gene.

INTRODUCTION

Studies on the β -globin gene complex in the mouse have demonstrated the existence of repeated DNA sequences interspersed throughout the intergenic regions (1, 2). These sequences are members of families of middle repetitive sequences and have been mapped to specific locations in the β -globin complex. Recently, seven of these repeated sequences have been shown to belong to the same family, specifically the L1Md family (3). This family of repetitive sequences is unusual in that the 3'-end of each element is conserved while truncation occurs at different positions of the 5'-end, resulting in variations in their length. Recently it has been proposed that the MIF-1, Bam 5, and R groups of repetitive sequences, previously thought to be independent of each other, are actually subcomponents of a single repeat called the BamHI family (4-8). Apparently, the L1Md family may also be grouped into this family (3).

The albumin and α -fetoprotein (AFP) genes of the mouse have been shown to be syntenic and are linked by a 13.5 kbp intergenic region (9, 10). In addition, this complex is approximately 50-55 kbp in size which is in the same size range as that of the β -globin complex (2). Since the L1Md family

is found dispersed throughout the mouse genome at a frequency of 10^4 times/genome, we undertook to determine if members of this family could be found in the albumin-AFP gene complex in the mouse. In these studies, we report on their localization in the mouse albumin-AFP gene complex at four specific sites, i.e. in intron 12 of the albumin gene, intron 3 of the AFP gene, and twice in the intergenic region. A preliminary report of these studies has appeared (11).

MATERIALS AND METHODS

Preparation of cDNA and genomic clones and subclones.

The mouse albumin (MSA1 and MSA3) and AFP (MAF1 and MAF2) genomic sequences, cloned into the EcoRI sites of Charon 4A, were a gift from Dr. James Bonner. These clones, isolated from a mouse plasmacytoma cell genomic library (12), contain sequences that cover the gene complex including sequences 2.8 kbp 5' to the albumin gene, 3.4 kbp 3' to the AFP gene, and all the intergenic region except for a 0.65 kbp segment.

The following repetitive sequence DNA probes (a gift from Dr. Marshall Edgell) were used to determine the location of the L1Md family of repeats: 1.35, a 1.35 kbp EcoRI restriction fragment lying between the β -major and β -minor genes of the Balb/c mouse. This fragment defines the 1.35 portion of the L1Md family (see Figure 1 and reference 3). X, is a 1.25 kbp EcoRI restriction fragment located 5' to the ϵ Y3 gene in the Balb/c mouse β -globin complex. This DNA sequence defines both the V and 1.35 portion of the L1Md family (Figure 1).

Isolation and purification of DNA.

DNA from the genomic clones was prepared according to the method of Enquist (13). DNA from pBR325 subclones was prepared by using a procedure from Norgard *et al.* (14).

Restriction analysis.

Restriction enzymes were obtained from Bethesda Research Laboratories, Gaithersburg, MD 20877 and used according to their specifications. In general, 1-2 μ g of DNA were digested with a 10-fold excess of restriction enzyme for 1 hr. at 37°C. The reaction was stopped by adding a half volume of 25% glycerol, 5% SDS, 0.4% bromophenol blue. Gel electrophoresis was carried out in 1.2% agarose in TA (40mM Tris, 5mM sodium acetate, 1mM EDTA, pH 7.8) at approximately 2 V/cm for 16 hrs. The DNA was then transferred to nitrocellulose filters.

Figure 1. A map of the consensus L1Md family of repeated sequences showing the probes used in these experiments and the regions these probes detect by hybridization. The region upstream from the 5'-end boundary of the U section is undefined (from reference 3).

Figure 2. EcoRI restriction map of the mouse albumin (A) and α -fetoprotein (B) genes showing the relative position of the genomic clones and EcoRI fragments subcloned into pBR325. Exons are represented by solid boxes while EcoRI sites by vertical lines. The EcoRI sites at the 5'-ends of fragments 1011 and 1-2 and the 3'-end of 3-3 were artificially generated during the cloning of the DNA fragments into Charon 4A. The 3'-end of the albumin map is contiguous to the 5'-end of the AFP map.

Transfer and hybridization.

DNA was blotted onto nitrocellulose filters using a modification of the method of Southern (15) published by Schleicher and Schuell in Technical Bulletin No. 352-354. The filters were prehybridized in 50% deionized formamide, 5x Denhardt's solution (1x is 0.02% each of ficoll, polyvinylpyrrolidone, and BSA), 5x SSPE (1x is 0.18M NaCl, 10mM NaPO₄, pH 7.7, 1mM EDTA), and sheared denatured salmon sperm DNA (150 µg/ml), allowing 50 µl of hybridization mix per cm² of filter. This was incubated at 42°C for at least 3 hrs. The filter is then hybridized in 50% deionized formamide, 1x Denhardt's solution, 5x SSPE, and 150 µg/ml of sheared denatured salmon sperm DNA. Between 5x10⁵ and 5x10⁶ cpm of denatured labeled probe was added. Hybridization was performed at 42°C for 36 hrs. The filter is washed four times at room temperature in 2x SSPE, 0.1% SDS; twice at 52°C in 0.1x SSPE, 0.1% SDS, air dried, and autoradiographed for approximately 24 hrs. using Kodak X-Omat film with a DuPont Chronex Lightening Plus intensifying screen at -90°C.

Hybridization probes were prepared by nick translation using [³²P]dCTP (Amersham Corp., Arlington Heights, IL 60005) and a kit supplied by Bethesda Research Laboratories. Before use, the probe is denatured in 50% deionized formamide at 95°C for 5 min. and then cooled in an ice bath.

Southern transfers and hybridizations carried out using Genescreen membrane (New England Nuclear, Boston, MA 02118) were done according to recommendations.

DNA sequencing.

Fragments of the repeated sequence localized in intron 12 of the albumin gene, bounded by the BamHI site at 15,850 and the PvuII site at 16,500, were cloned into M13mp10 and M13mp11 according to the procedure described in the M13 Cloning and Sequencing Handbook from Amersham Corp. The fragments cloned are from BamHI (15,850) to SstI (16,130) and SstI (16,130) to PvuII (16,500). Sequencing reactions were carried out according to the method of Sanger (16).

Figure 3. (A) Location of sequences homologous to X DNA (V + 1.35) in the mouse albumin-AFP gene complex. Subclones of the mouse albumin-AFP complex were digested with EcoRI and subjected to Southern blot analysis. The DNA fragments were hybridized with ³²P-labeled X DNA. Fragments 3-3, 1-3, 1-1, 1-4, and 1031 were found to have sequences homologous to X DNA.

(B) Location of sequences homologous to 1.35 DNA in the mouse albumin-AFP complex. ³²P-labeled 1.35 DNA was used for hybridization of Southern blots of a gel similar to the one described above. The autoradiogram shows that sequences homologous to the 1.35 DNA are only found in fragment 1-1.

RESULTS

Characterization of the lambda genomic clones and subclones.

Genomic clones (MSA1, MSA3, MAF1, MAF2) of the mouse albumin-AFP gene complex (Figure 2) were digested with EcoRI and each fragment was subcloned into the EcoRI site of pBR325 (17). These fragments were then used for preliminary localization of the L1Md repeated sequences throughout the complex.

Localization of the L1Md family of repeats in the albumin-AFP complex.

The X and 1.35 sequences define the repeated sequences found at several sites throughout the mouse β -globin gene complex (2). Our experiments were designed to determine if these sequences could be found throughout the albumin-AFP gene complex in the mouse. The subclones described in the previous section were digested with EcoRI and resolved by agarose gel electrophoresis. Southern analysis of this gel demonstrated that [32 P]a-labeled X sequence hybridizes to the 3'-half of the albumin gene (fragments 1-3 and 3-3), the 5'-half of the AFP gene (fragment 1031), and at two locations in the intergenic region (fragments 1-1 and 1-4) (Figure 3A). A similar experiment using [32 P]labeled 1.35 sequence as a probe showed that this DNA hybridizes to fragment 1-1 and thus is only found in one location throughout the complex (Figure 3B).

To further locate the position of these two sequences, additional restriction digests of the positive clones were examined by Southern analysis. The 1.35 sequence was localized to a 0.8 kbp fragment lying between the PvuII site at 24,350 and the EcoRI site at 25,150 (Figure 4). Additional experi-

Figure 4. A restriction map of the intergenic region between the albumin and AFP genes. The regions homologous to the 1.35 and X sequences are indicated by cross hatching and slanted lines respectively. The distances and sites of restriction endonuclease cleavage are expressed as base pairs starting at the 5'-end of the MSA-3 clone which extends 2.86 kbp upstream from the first exon of the albumin gene.

ments have revealed that the 1.35 sequence is located within the 100 bp between the HindIII site at 25,050 and the EcoRI site at 25,150 (Figure 4). The X sequence was localized to the same 0.8 kbp fragment as defined above and was shown to hybridize to either side of the HindIII site at 25,050. Thus, the 1.35 and X sequences hybridize to adjacent regions at this location.

Additional experiments were carried out to localize the X sequence in the albumin gene (fragment 1-3), in the intergenic region (fragment 1-4), and in the AFP gene (fragment 1031). The results of these experiments are summarized in Figures 4, 5 and 6. The X sequence is found in intron 12 of the albumin gene between the BamHI site at 14,990 and the PvuII site at 16,500 (Figure 5), in the intergenic region between the HindIII site at 27,950 and the EcoRI site at 28,850 (Figure 4), and in intron 3 of the AFP gene between the PvuII site at 36,200 and position 37,200 (Figure 6). Fragment 1011, which begins at this position, defines the latter coordinate as it does not hybridize to the probe (Figure 2). To establish the 3' margin of the repeat in intron 12 of the albumin gene (Figure 5), [³²P]labeled mouse genomic DNA was used to hybridize against a Southern blot of fragment 1-3 cut with several restriction enzymes. This probe should hybridize to any fragment that has a moderately repetitive sequence in it. The fragments which hybridize in the figure were found to correspond to the genomic sequence between the BamHI site at 14,990 and the PvuII site at 16,500. DNA on either side of this 1.5 kbp region does not contain a repetitive sequence as detected by this assay. Thus, the X sequence does not extend past the PvuII site at

Figure 5. A restriction map of the albumin gene in the region of exons 1 and m and intron 12. The position of the repeated sequence homologous to X DNA is indicated by the slanted lines.

Figure 6. A restriction map of the 5'-end of the mouse AFP gene in the region of exons c and d and intron 3. The position of the sequence homologous to the X DNA is indicated by the slanted lines.

16,500. Finally, since our clones of the intergenic region do not include approximately 650 bp downstream of the artificial EcoRI site at 28,850 (Figure 4), we cannot discount the possibility that the repeat may be longer at this location.

Sequence of the repeated sequence in intron 12 of the albumin gene.

To confirm the homology between the L1Md family of repeated sequences and those found in the albumin-AFP gene complex, sequence analysis was performed on the repeated sequence in intron 12 of the albumin gene. The 1.51 kbp fragment between the BamHI site at 15,850 and the PvuII site at 16,500 was cloned into M13 and sequenced by the method of Sanger (16). The sequence in Table I shows a high degree of homology with the consensus sequence of the 3'-end of the L1Md family (3). These data confirm the identity of the repeated sequences in the albumin-AFP complex and also establishes the orientation of the sequence in the albumin gene. It should be noted that by restriction mapping analyses, the 5'-end of this repeated sequence may extend to the BamHI site at 14,990. Nevertheless, our albumin cDNA clone does not hybridize to the X sequence, demonstrating that the repeat does not include exon 1. A more precise localization of this boundary has not been possible due to a paucity of restriction sites, although continued sequence analysis may yet determine this.

DISCUSSION

The existence of highly repetitive DNA sequences was first detected by DNA-DNA reassociation kinetics (18) and confirmed through their isolation and

TABLE I

A comparison of the nucleotide sequence of X DNA repeated sequence in intron 12 of the mouse albumin gene with the consensus nucleotide sequence of the R repeated sequence. The DNA fragment between the BamHI and PvuII site subcloned into M13 for sequencing. The top row (I12) represents the sequence determined for the albumin intron 12 fragment and the bottom row denotes the R family (Rcon). The blank spaces of Rcon denote where the two sequences match; the letters denote mismatches and insertions or deletions are denoted by a hyphen.

I12	GGATCCACCC	CATAAT-AGA	CATCCAAACG	CTGACACCAT	40
Rcon	T	C C -	C AN	T	
I12	TGCATACACT	AGCAAGATT	TATTGAAAGG	ACGCAGATGT	80
Rcon	TN T	AN	GC	NN T A	
I12	AGCTGTCTCT	-GTGAGACTA	TGCCGGGGCC	CAGCAAC---	120
Rcon	T		A N	T ATAC	
I12	AGAAGTGGAT	GCTCACGATC	AGTCATCTAG	TGGATGGATC	160
Rcon		----	N T	G	
I12	ATA--GGGCT	CCCAATG-AG	GAGCTAGAGA	AA-TAGCCAA	200
Rcon	C CA C	G		G C	
I12	GGAGCTAAAG	GGACCTTCAA	CCCTATAGGT	GGAACAACA-	240
Rcon		NT G		A	
I12	TAT--ACTAA	CCCATACCCC	GAGAGCTCTT	GACTCTAGCT	280
Rcon	GA	AG	C N T		
I12	GCATATATAT	CAAAGATGG	CCTAGTCGGA	CATCACTTGG	320
Rcon	G		C	G	
I12	AAGAGAGCCC	CATTGGACTT	GCAAA-TTTA	TATGCCCCAG	360
Rcon	G		T C		
I12	TACAGGGGAA	T-ACAGGGCC	AAAAAGGGGG	AGTGGGTGGG	400
Rcon		CGC	T		
I12	CA				402
Rcon	T				

sequencing (4, 8, 19-21). Extensive studies have established that these sequences can be grouped into several categories which include multigene families, satellite DNAs, short interspersed elements, and long interspersed elements (8, 21). The long interspersed repeats, designated LINES (22), are generally over 5 kbp in size and have been found dispersed throughout the genome at a repeated frequency of about 10^4 times. Repetitive sequences that fall into the category of the LINE sequences have been characterized in the mouse genome (4-7, 23). These sequences, MIF-1, Bam-5, and R are all found

in different abundance in the mouse genome leading to the assumption that they are all independent of each other. However, it has recently been proposed that each of these apparently independent families may be subcomponents of a larger repeated sequence or superfamily (3, 8). Thus, the MIF-1, Bam-5, and R groups are linearly arranged repeated sequences that comprise what has been defined as the L1Md family (3). In the map shown in Figure 1, therefore, the V segment which is at the 3'-end corresponds to the 3'-end of the Bam-5 family and includes the R family region; the 1.35 segment corresponds to the 5'-end of the Bam-5 and the U segment corresponds to the most 3'-end of the MIF-1 family (see figure 3 in reference 3).

Certain features of the L1Md family of repeats indicate that these elements may be propagated throughout the genome by an RNA intermediate. These sequences have been shown, for example, to possess poly-dA tracks and to be flanked by short direct repeats (8). In addition, the apparent conservation of the 3'-end in almost every element, while the 5'-end is found truncated at varying positions, is a characteristic of reverse transcription. This process may still be occurring in mice since RNA transcripts of this family are readily detected (7, 24).

This mechanism of propagation of the repeated sequences throughout the genome has been contested (20). Meunier-Rotival *et al.* (20) have pointed out that (a) truncation is not a common phenomenon in sequences 5' to the Bam-5 region of the superfamily, (b) that the linkage of MIF-1/Bam-5 (5') and R repeats (3') is unbalanced due to a shortage of R repeats, (c) the R repeats are much less abundant than the MIF-1/Bam-5 repeats, and (d) that some members of the superfamily extend past the poly-dA end of the R repeats. These conclusions were derived from studies of the distribution of the repeated sequence family in the entire genome. Results from studies with specific cloned repeated sequences support the mechanism based upon truncation. Since our studies agree with the mechanism based upon truncation, we hypothesize that this may account for the structure of the repeated sequences we observed in the albumin-AFP complex.

The L1Md family of repeated sequences has been found in seven locations in the 50 kbp mouse β -globin complex (3). Our studies have demonstrated that the L1Md family (or members of the L1Md family) are found twice in the intergenic region, once in intron 12 of the albumin gene, and once in intron 3 of the AFP gene. In three of these sites we find no evidence of the presence of fragment 1.35. Based on the concept that these sequences are linearly arranged, we conclude that at these 3 sites, termination (or truncation)

occurs at a point within the V region (Figure 1). On the other hand, we have also observed that one of the sequences located in the intergenic region includes the V sequences and part of the 1.35 segment. The X probe, which contains both V and 1.35 sequences, hybridizes to both sides of the HindIII sites at 25,050 while the 1.35 probe hybridizes only to the 100 bp 3' to this site. We conclude that this particular element of the L1Md family has been truncated within the 1.35 region.

We have also used the EcoRI U fragment from CE18 (1), which defines a segment of L1Md 5' to the 1.35 segment, and have detected no evidence indicating that this sequence is located in the albumin-AFP gene complex (data not shown). This further supports our hypothesis that the L1Md family of repeats in the complex is truncated within the V section in 3 out of 4 cases and within the 1.35 section in the fourth.

The elements of the L1Md family found in the albumin-AFP complex most closely resemble R family repeats as they are much shorter than the canonical repeat and, in general, only encompass the V segment. This was confirmed by the high degree of homology between the sequence in intron 12 of the albumin gene and the consensus sequence for the R repeat. Two of these elements are found within introns which may account for the length of these introns compared to the other introns in these genes. Larger insertions within an intron may not be viable, giving rise to the observation that only very short versions of the element are found at these locations.

We cannot at this time propose a function for these repeated sequences in this gene complex. As has been proposed, they may serve no function at all, and may be inserted at different places in the genome at random (25). On the other hand, we cannot discount the possibility that they may be involved in the control of these genes, possibly through their contribution to chromatin structure. The two complexes in the mouse where the L1Md family has been found, i.e. the β -globin complex and the albumin-AFP complex, are both developmentally regulated. A recent report has linked the transcription of an Alu family repeated sequence with the developmental expression of the ϵ -globin gene in mice (26). Although it is well known that repeated sequences are transcribed, their role in gene regulation remains to be clarified.

ACKNOWLEDGEMENTS

1. This investigation was supported by PHS grant CA30043 awarded by the National Cancer Institute, DHHS, to J.P.; by grant H-855 awarded by the Robert A. Welch Foundation to J.P.; by PHS grant CA17701-09 awarded by

the National Cancer Institute, DHHS, to the UTMB Cancer Center; by NSF grant SPI 8165065 awarded to W.F.Z.; and NIH grant 5P35NS07189 Short Term Training: Students of Professional Schools (C.E.H.).

2. We thank Dr. Marshall Edgell for providing DNA samples of the X, 1.35, and U sequences and Dr. James Bonner for the albumin and AFP clones.
3. We would like to thank Miss Mary C. Higgins and Mrs. Leanna Reynolds-Jackson for their assistance in the preparation of the manuscript.

*To whom reprint requests may be sent

REFERENCES

1. Jahn, C.L., Hutchison III, C.A., Phillips, S.J., Weaver, S., Haigwood, N.L., Voliva, C.F., and Edgell, M.H. (1980). *Cell* **21**, 159-168.
2. Haigwood, N.L., Jahn, C.L., Hutchison III, C.A., and Edgell, M.H. (1981). *Nucleic Acids Res.* **9**, 1133-1150.
3. Voliva, C.F., Jahn, C.L., Comer, M.B., Hutchison III, C.A., and Edgell, M.H. (1983). *Nucleic Acids Res.* **11**, 8847-8859.
4. Gebhard, W., Meitinger, T., Höcht, J., and Zachau, H.G. (1982). *J. Mol. Biol.* **157**, 453-457.
5. Meunier-Rotival, M., Soriano, P., Cuny, G., Strauss, F., and Bernardi, G. (1982). *Proc. Natl. Acad. Sci. USA* **79**, 355-359.
6. Brown, D.D.M. and Dover, G. (1981). *J. Mol. Biol.* **150**, 441-466.
7. Fanning, T.G. (1982). *Nucleic Acids Res.* **10**, 5003-5013.
8. Fanning, T.G. (1983). *Nucleic Acids Res.* **11**, 5073-5091.
9. D'Eustachio, P., Ingram, R.S., Tilghman, S.M., and Ruddle, F.H. (1981). *Somatic Cell Genet.* **7**, 289-294.
10. Ingram, R.S., Scott, R.W., and Tilghman, S.M. (1981). *J. Biol. Chem.* **256**, 4694-4698.
11. Papaconstantinou, J., Cooper, R., Herzog, C.E., and Li, M.-L. (1983). *Fed. Proc.* **42**, 2264.
12. Davis, M.M., Calame, K., Early, P.W., Livant, D.L., Joho, R., Weissman, I.L., and Hood, L. (1980). *Nature* **283**, 733-739.
13. Enquist, L.W., Madden, M.J., Schiop-Stansly, P., and Vande Woude, G.F. (1979). *Science* **203**, 541-544.
14. Norgard, M.V., Emigholz, K., and Monahan, J.J. (1979). *J. Bact.* **138**, 270-272.
15. Southern, E.M. (1975). *J. Mol. Biol.* **98**, 503-517.
16. Sanger, F., Nicklen, S., and Coulson, A.R. (1977). *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467.
17. Bolivar, F. (1978). *Gene* **4**, 121-136.
18. Britten, R.J., and Kohne, D.E. (1968). *Science* **161**, 529-540.
19. Wilson, R. and Storb, U. (1983). *Nucleic Acids Res.* **11**, 1803-1817.
20. Meunier-Rotival, M. and Bernardi, G. (1984). *Nucleic Acids Res.* **12**, 1593-1608.
21. Jelinek, W.R. and Schmid, C.W. (1982). *Ann Rev. Biochem.* **51**, 813-844.
22. Singer, M.F. (1982). *Cell* **28**, 433-434.
23. Heller, R. and Arnheim, N. (1980). *Nucleic Acids Res.* **8**, 5033-5042.
24. Soriano, P., Meunier-Rotival, M., and Bernardi, G. (1983). *Proc. Natl. Acad. Sci. USA* **80**, 1816-1820.
25. Orgel, L.E. and Crick, F.H.C. (1980). *Nature* **284**, 604-607.
26. Allan, M. and Paul J. (1984). *Nucleic Acids Res.* **12**, 1193-1200.