Volume 11 Number 16 1983 Nucleic Acids Research

Homology between the *Kpn*I primate and *Bam*H1 (M1F-1) rodent families of long interspersed repeated sequences

Maxine F.Singer, Ronald E.Thayer, Giovanna Grimaldi, Michael I.Lerman and Thomas G.Fanning+

Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205, and +Department of Pathology, School of Medicine, University of California, Davis, CA 95616, USA

Received 8 June 1983; Accepted 25 June 1983

ABSTRACT

The KpnI and BamHl (or M1F-1) families are the predominant sets of long interspersed repeated DNA sequences (LINEs) in primates and rodents, respectively. Recently, the sequences of several cloned subsegments from each family were determined in different laboratories. These sequences have now been compared and found to be homologous over at least 1400 bp. The data suggest that the two LINE families had a common progenitor and have been conserved in similar abundance although in divergent forms in the two mammalian orders.

INTRODUCTION

Families of highly repeated, interspersed sequences whose members can be 5 to 6 kbp or longer have been reported in primates (1-5) and in rodents (6-10) and called KpnI and BamHl (or M1F-1), respectively. Collectively, the families have been termed LINEs, for long interspersed sequences (11,12). Both occur on the order of 10^4 times in their respective genomes. Manuelidis and Biro reported that cloned subsegments of the KpnI LINE family hybridize weakly to restriction endonuclease digests of mouse DNA and that fragments of mouse LINEs hybridize to human DNA (3,13). Recently, Kole and coworkers confirmed the hybridization of KpnI family segments to mouse DNA (14). Others, including ourselves, were previously unable to detect such homology (5,10,15). However, we have now found weak cross-hybridization between two non-overlapping portions of KpnI family sequences and mouse DNA thereby confirming the positive experiments. Moreover, we find extensive homology between the primate KpnI and rodent BamHl (M1F-1) LINE families by direct comparison of primary nucleotide sequences.

METHODS

The sequence of E24 was determined as follows. Mouse genomic DNA was cleaved with endonuclease EcoRI and fragments of 2.5-3 kbp were isolated (15)

and cloned in the vector pUC9 (16). Inserts containing BAM5 sequences were isolated (15,17) and characterized by restriction mapping and sequencing (T.G. Fanning, accompanying manuscript). One plasmid, pEf1-2.4, was digested with endonucleases EcoRI and BamH1 and a 274 bp fragment abutting the BAM5 sequence was isolated. The fragment (E24) was recloned in M13mp8 and M13mp9 (18) and sequenced in both orientations by the dideoxy method (19,20).

Sequences were compared using computer programs that match sequences (21) or produce optimal sequence alignments (22).

RESULTS

Maps of KpnI and BamH1 (M1F-1) family members. Fig. 1A shows a schematic diagram of a long African green monkey (AGM) KpnI family member (5); it does not necessarily represent the structure of any particular family member since they are polymorphic regarding restriction endonuclease sites and length. The map is similar to that of typical human family members (3,4). Also indicated on Fig. 1A is the region (cross-hatch) homologous to the 1.9 kbp HindIII fragment of the human KpnI family, the sequence of which has been reported (23). Fig. 1B is the map of a short KpnI LINE family member called KpnI-LS1 (24). The precise length of KpnI-LS1 is unknown but it does not exceed 2.4 kbp; the sequence of 1789 bp of KpnI-LS1 is known as indicated by the cross-hatching in Fig. 1B. Fig. 1C shows a schematic diagram of a long BamH1 (M1F-1) family member based mainly on the work of Bernardi and coworkers (9,10); again, particular family members show restriction endonuclease site and other polymorphisms. The regions for which primary nucleotide sequence data are available are indicated by cross-hatching: M1FC37 (25), M1FC202 (26), E24 (this paper) and BAM5 (15,27). The alignment of the sequenced fragments with the detailed maps of typical BamH1 (M1F-1) family members was not described in the papers reporting the sequence data. We have deduced an alignment of the sequenced fragments by comparing the sequences and the restriction endonuclease sites contained therein with the published maps (8-10). The critical points in the alignment were 1) the positioning of the BglII site in the 1.3 kbp EcoRI fragment relative to the 0.5 kbp BamH1 (BAM5) fragment (9), and 2) the location of the HindIII site in the 1.3 kbp EcoRI fragment (9) that correlates with the HindIII site in the sequence of M1FC37 (25).

Cross hybridization of LINE family segments. Cloned probes from within AGM KpnI family members hybridize weakly to restriction endonuclease digests of total mouse DNA (not shown). For these experiments we used the two subclones

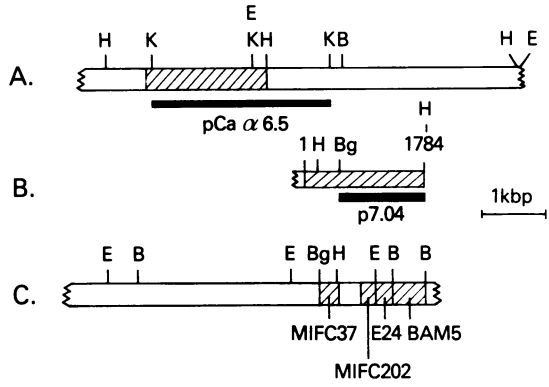

Fig. 1. Maps of monkey KpnI and mouse BamHl (MlF-1) family members. Cross-hatches indicate regions of known sequence. A. A composite of a typical long AGM KpnI segment (5). The sequenced region is a 1.9 kbp HindIII fragment of human DNA (23); it is placed within the typical AGM 2.5 kbp HindIII fragment on the basis of restriction endonuclease mapping and cross-hybridization (5). The bar below indicates the monkey segment included in the subclone pCaα6.5 (5). B. The AGM KpnI-LS1 segment whose partial sequence (1784 bp) is shown in Fig. 2 (residues 1 and 1784 are indicated here) (24). The bar below indicates the segment included in the subclone p7.04 (24). C. A composite of a mouse BamHl (MlF-1) segment (8-10). The sequenced regions are: MlFC37, a 238 bp mouse segment (25); MlFC202, a 250 bp mouse segment (26); E24, a 274 bp fragment whose sequence is reported in Fig. 2; BAM5, a 507 bp mouse segment (15,27). Restriction endonuclease sites are: B, BamHl; Bg, BglII; E, EcoRI; H, HindIII; K, KpnI. Only a few of the known endonuclease sites are shown. In B and C the segments are aligned with the homologous regions in A (see Fig. 2 and reference 24).

indicated on Fig. 1A and 1B. The pCaα6.5 probe hybridized to a 4 kbp BamHl fragment of mouse DNA as well as to a smear of fragments of various sizes; it also hybridized to a smear of HindIII fragments. The p7.04 probe hybridized to a 5.5 kbp MspI band as well as a smear of MspI fragments. Both a 4 kbp BamHl (9,10) and a 5.4 kbp MspI fragment (15) are produced from the BamHl (MlF-1) mouse LINE family; the 4 kbp BamHl fragment is shown on Fig. 1C but the position of the 5.4 kbp MspI fragment is unknown. Thus two non-overlapping portions of KpnI family segments hybridize weakly to genomic mouse bands of a size consistent with their being part of the BamHl (MlF-1) family.

Comparison of nucleotide sequences. In order to establish the significance of the cross-hybridization, we compared the available KpnI and BamHl (MlF-1) family sequences (see Fig. 1). All the sequence data were derived from single cloned DNA fragments and therefore they probably diverge to an unknown extent from the consensus sequences for the two families. No significant long re-

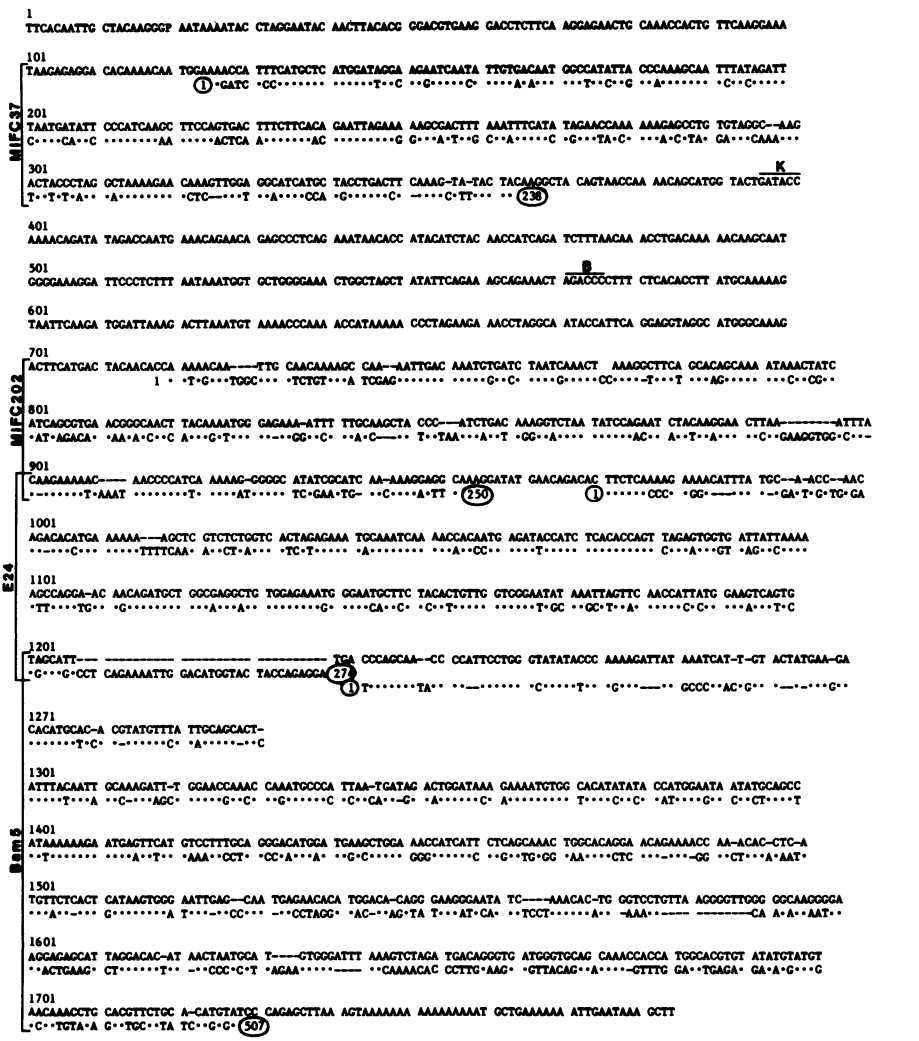

Fig. 2. Sequence comparison of KpnI-LS1 with BamH1 (M1F-1) segments. The 1784 sequenced base pairs of KpnI-LS1 (24) are on the upper line numbered as in the initial report. On the lower line are the mouse sequences M1FC37 (25), M1FC202 (26), E24 and BAM5 (15); the residue numbers used in the original publications are encircled at the beginning and end of each. Residues in the mouse sequences that are identical to KpnI-LS1 are shown as a dot; hyphens indicate deletions in one or the other sequence. All sequences are 5' to 3', left to right and top to bottom. The region corresponding to each mouse segment is indicated by the side brackets. K and B mark, respectively, the positions of KpnI and BamH1 sites typical of many AGM KpnI family members (see Fig. 1A); both sites are diverged in KpnI-LS1.

gions of homology were found between the human 1.9 kbp HindIII fragment (Fig. 1A) and any of the known mouse sequences. However extensive homology was found between the 1784 bp of the AGM KpnI-LS1 sequence and each of the four known mouse sequences. The regions of homology are shown on Fig. 2. The entire reported sequence of M1FC37 (238 bp) lines up with a continuous stretch of KpnI-LS1 with 69 percent homology. Then, after a 357 bp gap, the entire reported sequence of M1FC202 (250 bp) aligns with 60 percent homology. Thereafter, 241 bp of the 274 bp in the E24 fragment described in this report align with a homology of 70 percent; the final 33 bp of E24 do not match the contiguous segment of KpnI-LS1. According to the BamH1 (M1F-1) family map (Fig. 1C), the end of E24 (residue 274) and the beginning of BAM5 (residue 1) should be within the same BamH1 site. Recently, Wilson and Storb (27) sequenced across this boundary in a cloned segment (CλB); comparison of the CλB sequence with those of E24 and BAM5 confirm the contiguity of the latter two segments. Therefore it appears that BamH1 (M1F-1) family members may contain about 30 bp (residues 242-274 of E24) that do not appear in KpnI-LS1. Finally, the first 311 bp of BAM5 line up with a 67 percent homology to KpnI-LS1; the remaining 199 bp of BAM5 have less than 50 percent homology to KpnI-LS1.

Comparison of the maps in Fig. 1 with the sequences in Fig. 2 suggest that the homologous regions in KpnI-LS1 and BamH1 (M1F-1) family members are colinear over at least about 1.4 kbp. This conclusion assumes that the homology will extend into the present gap in the known BamH1 (M1F-1) sequence. In this regard we note that the estimated 360 bp of undetermined sequence in the 850 bp BglII/EcoRI BamH1 (M1F-1) family segment about equals the length of the present 357 bp gap (residues 363-719 on KpnI-LS1).

The homology revealed by the sequence comparison is consistent with the fact that pCaα6.5 hybridized with the 4 kbp BamH1 fragment. In their experiments, Manuelidis and Biro (3,13) used the human 1.9 kbp HindIII fragment (see Fig. 1A) as a probe; since all of the sequences in this probe are outside the sequenced regions shown in Fig. 2, it appears that homology between the two families extends well beyond the 1.4 kbp indicated by our data.

DISCUSSION

Our findings indicate that the approximately 10^4 interspersed copies of BamH1 (M1F-1) and KpnI LINEs in rodent and primate genomes are related. The relation is distant as indicated by the distinct typical restriction endonuclease maps and by the 30-40 percent sequence divergence in those regions

that have been compared. It is not possible to give a precise number for the sequence homology since individual family members diverge within a species and no consensus sequences are available.

We conclude that the two orders of mammals have distinctive versions of related interspersed long repeated DNA families just as they have distinctive versions of related interspersed short repeated families, i.e., Alu and B1 (12,28). Within each order there are species specific differences in LINE families; restriction endonuclease polymorphisms differentiate most of the human from most of the monkey KpnI family members (2,3,5) and there is also a variable abundance of divergent BamH1 (M1F-1) subfamilies among species of rodents (8). The data suggest that the LINE families had their origin in a common progenitor before the separation of the rodent and primate lineages some 70 ± 5 x 10^6 years ago and further that within each species, the LINE family members have undergone concerted evolution (29-31). The rate of divergence suggested by our comparison (about 0.5 percent per million years) indicates that the rodent and primate families are diverging at a rate slightly lower than that reported for divergence in introns and silent coding positions (32). The divergence of the second unit of the dimeric Alu consensus sequence from the homologous regions of the B1 consensus is 23 percent, somewhat less than the observed divergence between the LINE families. However, these points must be considered tentative until consensus sequences rather than individual cloned sequences can be compared. Similarly, the significance of the marked divergence between KpnI-LS1 and the last 200 bp of the BAM5 sequence must await additional analysis.

The function of LINEs is now unknown. Recently, RNAs homologous to BamH1 (M1F-1) and KpnI family sequences were detected in mouse (10,15) (S.-M. Cheng, personal communication) and monkey and human (14,24) cells, showing that at least some family members are transcribed. Characterization of the transcripts may eventually contribute to an understanding of the functional role(s) of LINE families and of why so many family members are maintained in the genomes.

ACKNOWLEDGMENTS

We thank Giorgio Bernardi, Stephen D.M. Brown and S.-M. Cheng for communicating their data prior to publication and E. Kuff, T. Schopf, G. Heidecker and R. Cardiff for discussion. We are also grateful to May Liu who prepared the manuscript. One of us (T.G.F.) was supported by PHS Grant 5 RO1CA21454 from the NIH.

REFERENCES

1. Adams, J.W., Kaufman, R.E., Kretschmer, P.J., Harrison, M. and Nienhuis, A.W. (1980) Nucleic Acids Res. 8, 6113-6128.
2. Shafit-Zagardo, B., Maio, J.J. and Brown, F.L. (1982) Nucleic Acids Res. 10, 3175-3193.
3. Manuelidis, L. and Biro, P. (1982) Nucleic Acids Res. 10, 3221-3239.
4. Shafit-Zagardo, B., Brown, F.L., Maio, J.J. and Adams, J.W. (1982) Gene 20, 397-407.
5. Grimaldi, G. and Singer, M.F. (1983) Nucleic Acids Res. 11, 321-338.
6. Horz, W., Hess, I. and Zachau, H.G. (1974) Eur. J. Biochem. 45, 501-510.
7. Cheng, S.-M. and Schildkraut, C.L. (1980) Nucleic Acids Res. 8, 4077-4089.
8. Brown, S.D.M. and Dover, G. (1981) J. Mol. Biol. 150, 441-466.
9. Meunier-Rotival, M., Soriano, P., Cuny, G., Strauss, F. and Bernardi, G. (1982) Proc. Natl. Acad. Sci. U.S.A. 79, 355-359.
10. Soriano, P., Meunier-Rotival, M. and Bernardi, G. (1983) Proc. Natl. Acad. Sci. U.S.A. 80, 1816-1820.
11. Singer, M.F. (1982) Cell 28, 433-434.
12. Singer, M.F. (1982) Int. Rev. Cytol. 76, 67-112.
13. Manuelidis, L. (1982) In Genome Evolution (eds. Dover, G.A. and Flavell, R.B.) Academic Press, 263-285.
14. Kole, L.B., Haynes, S.R. and Jelinek, W.R. (1983) J. Mol. Biol. 165, 257-286.
15. Fanning, T. (1982) Nucleic Acids Res. 10, 5003-5013.
16. Vieira, L. & Messing, J. (1982) Gene 19, 259-268.
17. Hanahan, D. and Meselson, M. (1980) Gene 10, 63-67.
18. Messing, J. and Vieira, J. (1982) Gene 19, 269-276.
19. Sanger, F., Nicklen, S. and Coulsen, A. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
20. Heidecker, G., Messing, J. and Gronenborn, B. (1980) Gene 10, 69-73.
21. Queen, C. and Korn, L. (1980) Methods Enzymol. 65, 595-609.
22. Wilbur, J. and Lipman, D. (1983) Proc. Natl. Acad. Sci. U.S.A. 80, 726-730.
23. Manuelidis, L. (1982) Nucleic Acids Res. 10, 3211-3219.
24. Lerman, M., Thayer, R.E. and Singer, M.F. (1983) Proc. Natl. Acad. Sci. U.S.A., in press.
25. Brown, S.D.M. and Piechaczyk, M. (1983) J. Mol. Biol. 165, 249-256.
26. Brown, S.D.M. (1983) Gene, in press.
27. Wilson, R. and Storb, U. (1983) Nucleic Acids Res. 11, 1803-1817.
28. Jelinek, W.R. and Schmid, C.W. (1982) Annu. Rev. Biochem. 51, 813-844.
29. Zimmer, E.A., Martin, S.L., Beverly, S.M., Kan, Y.W. and Wilson, A.C. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 2158-2162.
30. Jeffries, A.J. (1982) In Genome Evolution (eds. Dover, G.A. and Flavell, R.B.) Academic Press, 157-176.
31. Dover, G., Brown, S., Coen, E., Dallas, J., Strachan, T. and Trick, M. (1982) In Genome Evolution (eds. Dover, G.A. and Flavell, R.B.) Academic Press, 343-372.
32. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980) Cell 20, 555-566.