Comparison of genomic fragment and clone sequences within a long interspersed repeated sequence of the mouse genome

Douglas L. Vizard and Julianne Yarsa

Department of Physics, University of Texas M.D. Anderson Hospital and Tumor Institute,
Houston, TX 77030, USA

Received 17 September 1984; Revised and Accepted 18 December 1984

ABSTRACT

The 393bp nucleotide sequence of a HindIII genomic fragment mapping within the major long interspersed repeated sequence family (MIF-1, Bam, Ll) of mouse is reported and compared to clone sequences of the same region of this repeated sequence. The consensus of the clone sequences significantly differs from the genomic fragment sequence by additions and deletions that are inconsistent with the physical and biochemical properties of the genomic fragment. While alternative explanations could account for some of these differences, several aspects of the experimental results imply that cloning artifacts contribute to the discrepancies. Despite the differences between the clone and genomic fragment sequences, the biologically interesting features previously noted in clone sequences (promoter-like signals and an open reading frame) are conserved in the genomic fragment sequence.

INTRODUCTION

The major long interspersed repeated nucleotide sequences (LINEs) of mouse and man have been shown to be distantly related (1), and speculations concerning their divergence and concerted evolution have recently been published (2,3). The rodent family of sequences has variously been called MIF-1, Bam and Ll; we shall call it Bam in this communication. Clone sequences of the Bam family have numerous differences; in particular, addition/deletions and base substitutions are frequent in clone comparisons (4). In the terms of concerted evolution, a consensus sequence of the cloned examples should reflect the current status of homogenization of a progenitor sequence if a dominant prototype exists. Both the narrow restriction bands observed by many investigators and the relative S1-resistance of Bam family heteroduplexes (5) indicate a dominant prototypic sequence. It follows that direct sequencing of selected genomic fragments from the Bam family may be possible. Furthermore, it is of interest to determine if the consensus sequence of the several cloned examples is converging to the genomic fragment sequence.

We have previously shown that the 400bp HindIII genomic fragment of mouse can be mapped into the Bam family LINE and can be prepared for sequencing (5). The sequence of this genomic fragment is of particular

interest, since clone sequences of this region of the Bam family possess promoter-like signals (6) and an open reading frame (7). The chemical consensus sequencing results are presented and are compared to clone sequences of the same region. The results imply that the consensus sequence of clones significantly deviates from the genomic fragment sequence. Additional data concerning restriction site and length polymorphisms of cloned fragments suggest that the observed disparities in the nucleotide sequences may be due to cloning artifacts generated by replication errors near symmetric sequences. However, the promoter-like signals and an open reading frame are conserved in the genomic consensus sequence.

MATERIALS AND METHODS

DNA sequencing reactions were performed according to a chemical methodology adapted from Maxam and Gilbert (1980), except that the thymine-specific cleavage was that of Rubin and Schmid (1980). Sequencing gel electrophoresis was done at high temperature in 50% urea, in gels whose acrylamide/bis ratios were 10/1, and in 50 mM Tris-borate (1 mM EDTA). The same buffer and room temperature were used for other gels, except for refrigerated strand-separation gels.

Blotting, probe hybridization and enzymatic processing of DNA were described previously (5). The limited quantities of genomic fragments necessitated optimizing fragment repurification protocols with respect to product yield and purity; we found that electro-elution of gel slices into hydroxylapatite and subsequent cationic detergent precipitation of the eluted fragment (10) gave the best results. Cloning of the 270bp EcoRI, BamHI mouse genomic fragment was done by ligation into an EcoRI, BamHI and phosphatased derivative of PBR322 (5), and transfection into E.coli ED8767 (11). The clones were selected by colony-lift hybridization.

RESULTS AND DISCUSSION

Because of its good resolution on gels, the HindIII 400 bp fragment generated by whole genomic digests presents an opportunity to directly sequence a region within the Bam family repeated element (5). Similarly, an EcoRI, BamHI digest of mouse DNA yields a clearly resolved 270bp fragment that overlaps the HindIII 400bp fragment in this region of the repetitive element (map in Fig. 1). Comparison of the overlapping sequences will address the possibility that the two fragments originate from differing subsets of the Bam family. From our blot hybridization studies (5 and unpublished), the EcoRI and BamHI sites occur in the vast majority of Bam family repeats, whereas the HindIII sites occur less frequently, possibly defining a subset. If distinct subsets of the Bam family are defined by different sequences in this region of the repeated element, it is likely that the nucleotide

Fig. 1. Preparation and sequencing strategy of genomic fragments. Photographic panels are autoradiograms of gels resolving the HinfI-cut and end-labeled genomic fragments: The HindIII 400bp preparation (left) was resolved in 8%(1/20 bis-)acrylamide and the repurified HinfI fragments 1,2, and 3 were strand-separated in 8%(1/80 bis-)acrylamide, where the lane numbers correspond to the HinfI fragment numbers. The EcoRI, BamHI 270bp preparation (right) was resolved in 8%(1/80 bis-)acrylamide, where the left and the right lanes are native and denatured material, and the numbers designate uncut (0), the 180bp (1) and the 97bp (2) fragments. The densitometer trace at the lower left is from an ethidium-stained preparative gel (3%(1/20 bis-)acrylamide) of EcoRI, BamHI digested C3H mouse DNA. The restriction map at the lower right shows how the HindIII 400bp and the EcoRI, BamHI 270bp genomic fragments map into the Bam LINE; the break-points in the map lie in the EcoRI 1.35kbp and BamHI 550bp regions of the Bam LINE. The genomic fragment sequencing strategy is shown by the arrows below the map.

sequence of these two fragments will differ.

Both the 400bp and 270bp genomic fragment preparations can be cut nearly quantitatively with HinfI as shown in Fig. 1. The same figure shows that the HinfI-generated fragments can be strand-separated, except for the 180bp HinfI, BamHI portion of the 270bp genomic fragment. The narrow bands generated by HinfI cutting imply homogeneity of duplex fragment lengths, whereas the broad bands in the strand-separation gel imply base sequence heterogeneity in these HinfI fragments, especially in fragment 3 of the HinfI-cut HindIII 400bp genomic fragment.

Since all fragments generated by these enzymes have 5' protruding ends, HinfI cutting, 5' end-labeling and strand-separation were used to prepare labeled single strands for chemical consensus sequencing according to the sequencing strategy shown in Fig. 1. Genomic fragments from stained or autoradiogrammed gel bands were inclusively excised to maintain yield. In no case was cross-contamination apparent in the sequencing gels, and such

Fig. 2. An example of a chemical sequencing gel autoradiogram, showing mapped residues from position 219 (bottom) to 266. In this gel, position 219 is 34 bases from the labeled end of the fast strand of the 97bp *Hinf*I fragment of the 270bp *Eco*RI, *Bam*HI genomic fragment preparation. The gel lanes (left to right) represent guanine, purine, cytosine and thymine specific reactions; the thymine lane has been differentially exposed to partially compensate for under-loading. Base assignments for this particular gel are shown at the right, where open circles designate ambiguity. The gel was 8% acrylamide.

liberal excisions precluded the possibility that fragment length homogeneity was experimentally introduced.

Although the bands of the strand-separated *Hinf*I fragments are broad, a clear consensus sequence of all strand-separable fragments can be read with few ambiguous assignments (N's). Figure 2 shows that the clarity of our sequencing gels is not a strong function of distance from the end-label, showing that the nature of the sequence heterogeneity in these genomic repeats must be that of base substitution rather than addition/deletion. For example, if addition/deletions among genomic copies were distributed at random (poisson) with a frequency of 4% per nucleotide, then only 37% of the signal associated with a residue 25 bases from an end-label would migrate at that position, another 37% would be divided between the 24 and 26 base positions, 18.5% between the 23 and 27 base positions, etc. On the basis of the guanine (the most specific chemistry) lanes of our sequencing gels, we can conservatively estimate that the frequency of addition/deletions in the genomic fragments is less than 0.2%. This conclusion is supported by the narrow double-stranded fragment bands compared to the broader bands in the strand-separating gels (Fig. 1).

The sequence for the *Hind*III fragment is given in Fig. 3. This 393bp genomic consensus sequence (H393CN) was derived from sequencing both strands

Fig. 3. The numbers and upper lines correspond to the HindIII 393bp genomic consensus sequence of C3H mouse. LEFT: Restriction sites related to text discussion are boxed and inverted repeated symmetries are designated by arrows. The lower lines compare the clone consensus sequence taken from Meunier-Rotival and Bernardi (1984) to H393CN, where the second line shows base substitutions and deletions (hyphens) and the third line shows insertions. The fourth line indicates evidence for deletions (-) and additions (+) from intercomparison of the clone sequences. RIGHT: Signals that potentially relate to expression are boxed and a stop codon is underlined. The lower lines compare the colinear portion of clone LIMd-4 from Martin et al. (1984) to H393CN using the same notations as in the left sequence. The genomic consensus sequence is that of the fast strands; all sequences are given 5' to 3'; N designates positions having little nucleotide preference.

of the strand-separated HinfI fragments. The sequence is that of the fast strands, and it is presented in the same orientation as the map in Fig. 1 (opposite to the orientation of Fig. 5 in ref. 5). The 97 base sequence of the EcoRI, HinfI element of the 270bp genomic fragment agreed perfectly with that of the overlapping portion of HindIII, HinfI fragment 1, except for a few positions where a majority nucleotide (rather than an N) was revealed from the comparison of sequences. The sequence agreement between genomic fragments terminated by different restriction sites shows that a repeated sequence subset was not selected on the basis of paired restriction sites. Therefore, this sequence should accurately reflect the "dominant prototype" among the genomic copies. Furthermore, it is not likely that sequence subsets of the Bam family are distinguishable in this region of the repeated element.

The chemical method of sequencing was chosen to obtain methylation information. Despite our efforts, C-5 methylation could not be clearly distinguished from an enhanced level of base substitution. We can only suspect that, for instance, the dinucleotide NG at position 56 is methylated at C5, but many other CG dinucleotides are frequently unmethylated (positions 225 and 233, see Fig. 2).

Clone and Genomic Sequence Comparisons. Our H393CN genomic consensus sequence is compared in Fig. 3 to a clone consensus sequence assembled by Meunier-Rotival and Bernardi (1984). Most of these clones were ascertained on the basis of the EcoRI site, and thus should be comparable to our genomic fragment sequence. The comparison shows 30 base substitutions (ignoring all Ns), 2 deletions and 17 additions (5 single and 6 double base insertions) with respect to H393CN. Most of the substitutions are transitions. Of the 24 Ns in the clone consensus and the 10 Ns in the genomic fragment sequence, only one corresponds in position; therefore, with respect to base substitution, the two consensus sequences deviate in 30 to 62 (8 to 16%) widely scattered positions.

Clearly, there is a great deal of base substitution among clones (4), but the lack of positional correspondence of the Ns between the clone consensus and genomic sequences suggests that the two sequences are not in the expected agreement. It is difficult to assess the significance of these differences, since the clone consensus sequence is taken from as few as two examples in some regions. However, the overall frequency of base substitution among clones is in approximate agreement with the average breadth of the bands in the strand-separation gels of the genomic fragments.

More significant discrepancies between the clone and genomic sequences are apparent in an analysis of their addition/deletion differences. Comparison among clones reveals numerous points where a minority of the compared clones show an addition or deletion with respect to the clone consensus sequence (4). At such positions, an appropriate + or - is noted in

Fig. 3. Of the 19 total addition/deletion deviations of the clone consensus from the genomic fragment sequence, 10 are coincident in position with addition/deletion inconsistencies (+/-) among clones. These coincident positions may be associated either with infrequent copies of the repeated sequence or may be examples of cloning artifacts, either of which might be generated by replication errors. It is noteworthy that 9 of the 10 coincident positions are in regions that may be subject to such errors, as discussed below.

There are 6 positions where double insertions appear in the clone consensus with respect to the genomic sequence. Five of those six positions are within or close to regions of inverted repeated sequence symmetry; the most significant of such symmetries are shown in Fig. 3. Other sequence symmetries such as direct repeats and palindromes exist at a significant level in H393CN, but are not denoted. Furthermore, 15 of the 19 total addition/deletion deviations between the clone consensus and genomic sequences are in these same regions. Such a coincidence correlates with the observation that cloning (12) and replication (reviewed in 13) artifacts have been observed in such regions of symmetry.

Numerous addition/deletion differences among copies of this repeated sequence are inconsistent with the fact that the genomic fragment sequencing gels are quite readable. On the basis of the addition/deletion frequency among clone variants (about 4%), few of our sequencing gels should have been readable beyond about 25 bases from the end-label. Our gels were generally legible through 100 nucleotides from the labeled end, being limited to that number by our necessary use of undried gels (0.8mm, frozen) and fluorescent enhancement screens. It is possible that a readable genomic consensus sequence could arise if additions and deletions were non-randomly distributed within individual genomic copies, where an addition is regionally paired with a deletion. Such would require a biological selection conserving the Bam family repeat length, which we have previously suggested (5). However, additional experiments, presented below, imply that the addition/deletions are a property of the clones as opposed to the genomic fragments, suggesting that cloning artifacts contribute to this apparent inconsistency.

Variation among Clones. An important observation suggesting the presence of cloning artifacts is the reduced recovery of restriction fragment sites in these clones. It is clear from the data in Fig. 1 that the genomic fragments are cut nearly quantitatively at positions 131 and 281 with HinfI. Probed blots of appropriately cut whole genomic DNA (not shown) confirm this observation. Other experiments show that the HindIII 400bp genomic fragment is quantitatively cut with EcoRI at position 184 and cut with about 70% efficiency with HaeIII at position 43. A study of the restriction enzyme recognition sequences among all clones referred to in Fig. 3 reveals that 1/4

Fig. 4. Restriction fragments of selected clones of the HindIII 400bp genomic fragment, where clone F is pH400 (5). Upper panel shows an autoradiogram of a gel blot probed with the excised insert of pH400, where the purified clones were cut with HinfI and HindIII. Lane ΔF is the same as lane F after reheating the probed blot to 48° in hybridization buffer (5). The lower panel is an ethidium stained gel of EcoRI cut clones, where one of the sites is in the plasmid. The letters correspond to many of the same clones in the upper panel. Lane M is a marker (HaeIII PM2). The insert of clone I has no EcoRI site; clone H is a double insert, only one of which has an EcoRI site. Arrows indicate the orientation of the inserts with respect to the usual PBR322 map. Both gels were 5%(1/20 bis-)acrylamide.

On the right are densitometer traces of autoradiograms from the probed blot at the upper left; the trace labels correspond to the lane labels.

had the HaeIII site at 43, 0/6 had the HinfI site at 131, 6/6 had the EcoRI site at 184, and 3/3 had the HinfI site at 281. Clearly, the HinfI site discrepancy at position 131 between the clones and the genomic fragment is a strong indication that the clones do not resemble the genomic fragment in this detail.

Our own clones of the HindIII 400bp fragment show a similarly poor HinfI site recovery at position 131. As shown in the probed gel blot of Fig. 4, only 3/6 of the clones studied had both HinfI sites; 6/6 had the HinfI site at 281. In 8/10 inserts, our clones possess the expected EcoRI site at position 184, as shown in the stained gel of Fig. 4, which also shows that the inserts are oriented without preference.

The more intensely homologous and smallest HinfI fragment 3 has a variable electrophoretic mobility among clones, as is apparent in Fig. 4. To study the electrophoretic mobilities, densitometer traces of probed blot autoradiograms of selected clones possessing both HinfI sites were compared,

as shown in Fig. 4. In this particular comparison (traces D and F), the measured variation of HinfI fragment 2 corresponds to 6 bp, and approximately 2 and 0 bp shifts are measured for fragments 3 and 1, respectively. These measured differences are typical for the HindIII, HinfI cut clones, although fragment 3 shows more electrophoretic variation among other clones than this example indicates. The mobility differences of the duplex fragments give additional evidence that addition/deletion is common among the cloned variants of this DNA sequence. However, such mobility differences are not consistent with the sharp bands found from HinfI digestion of the genomic HindIII fragment (Fig. 1). Hence, both a low recovery of restriction sites and addition/deletion variations among clones are apparent in our studies and are inconsistent with the corresponding properties of the genomic fragments.

Melting experiments were performed on the probed blot of Fig. 4 to determine if any of the three HinfI fragments in clones that possessed them demonstrate an unusually high degree of sequence divergence as measured by the loss of hybrid. An example of the melting experiment results is shown in autoradiogram traces F and ΔF in Fig. 4. In this case, the probed blot was heated to within a few degrees of the measured hybridized probe T_m , removing 65% and 80% of the hybrid to fragments 1 and 2 (respectively) compared to the fragment 3 hybrid. The genomic HindIII, HinfI fragments react similarly to clone F in this experiment (relatively inefficient hybridization to fragment 2), explaining the observation (5) that pH400 hybridizes with reduced efficiency to the EcoRI 1.35kbp genomic fragment. The relative order of hybrid loss is in accord with the respective G+C composition of the fragments. Hence, these data support the observation that the numerous differences among the repeated copies (mostly base substitutions) are widely scattered.

Since the above results indicate that the physical and biochemical properties of cloned fragments contrast with the properties of the genomic fragment preparations, we must conclude that some step in the molecular cloning of the genomic fragments can introduce the apparent addition/deletion artifacts. We have noted a rough correlation between addition/deletions and regions of inverted repeated sequence symmetry. Since these symmetries may be associated with replication errors (13), it is likely that different hosts might demonstrate different frequencies of such errors. To investigate this possibility, we cloned a series of fragments from an EcoRI, BamHI 270bp genomic preparation into the same vector (a derivative of pBR322) used for the HindIII 400bp fragment cloning (5). The host used in this experiment was E.coli ED8767, a well-characterized rec⁻ strain (11); the host used in the cloning of the HindIII 400bp fragment was a strain of E.coli HB101, whose rec⁻ status was verified. The two cloning experiments should be comparable to the extent that 76% of the 270bp EcoRI, BamHI fragment coincides with the

Fig. 5. Insert characterization of clones of the EcoRI, BamHI 270bp genomic fragment. The EcoRI, BamHI cut clones were resolved in 5%(1/20 bis-)acrylamide and stained with ethidium. Six inserts (lanes 1-6) are compared in the gel, where lane P is the pooled digests of all six clones. The 5bp deviation apparent in lane 1 is confirmed in lane P.

400bp HindIII fragment, and the coincident region lies in the two HinfI fragments (1 and 3) that show the most sequence heterogeneity in the strand separation gels (Fig. 1); HinfI fragment 3 also shows the greatest variation of duplex mobility among clones (Fig. 4). The inserts of six hybridization-positive clones propagated in *E.coli* ED8767 are characterized in the stained gel of Fig. 5, where only one in six clones had an insert that appreciably deviated (by 5bp) from the rest of the clone inserts. These results are in marked contrast to the length variation characteristic of the HinfI fragments of the HindIII 400bp fragment clones in Fig. 4, which were propagated in *E.coli* HB101. These contrasting results suggest that the apparent cloning artifact is a host-dependent function.

Biological Significance. Biologically significant aspects of this region of the Bam LINE have been discussed in other reports, where a PolII-like promoter structure (6) and an open reading frame (7) have been noted in cloned examples. As shown in Fig. 3, the PolII-like signals are indeed a feature of the genomic consensus sequence and an analysis of the sequence reveals a single open reading frame, excepting an opal codon at position 308. We note that our base assignment at position 309 is weak, but it is a marginal G. The CCAAT and TATA boxes are appropriately oriented with respect to the open reading frame and, remarkably, the five ATG and one ATN codons (potential translation starts) are all located downstream of TATA, and are all in the same open reading frame. In contrast to these functional correlates, we find that the normalized dinucleotide frequency within H393CN is intermediate between coding and non-coding sequences (14), and does not appreciably differ before and after the first ATG.

Figure 3 shows the comparison of the clone L1Md-4 to H393CN; it was on the basis of this clone sequence that Martin et al. (1984) reported the open reading frame. Alignment of L1Md-4 to H393CN produces 31 base substitutions,

7 additions and 3 deletions, showing somewhat less deviation from H393CN than does the clone consensus sequence in the same figure. The open reading frame of L1Md-4 is coincident with that of H393CN for 61% of the sequence, being phased in and out seven times by the addition/deletions in the sequence.

CONCLUSION

The genomic consensus sequence of a 393bp region of the Bam LINE of C3H mouse significantly differs from a consensus sequence derived from available clone sequences of the same region. The number of differences marked by base substitution is within reason, some of which may be an artifact of our judgmental assignment of a majority nucleotide and others may be due to the statistically low number of clones composing the consensus. The numerous addition/deletion differences among clones and between the clone consensus and genomic fragment sequences are contrary to the physical and biochemical properties of the genomic fragments, and are inconsistent with the readability of the genomic fragment sequencing gels. These discrepancies imply that cloning artifacts may be contributing to the observed differences. Characterization of our own clones confirms the aberrant nature of clones of this region of the Bam LINE and an additional experiment suggests that the cloning artifacts are host-mediated. Analysis of the sequencing data suggests that the focal points of cloning artifacts are in regions of inverted-repeated sequence symmetry. Although we believe that cloning artifacts contribute to the observed inconsistencies in the consensus sequences, it is possible that some of these discrepancies may derive from differing mouse lines and strains that have been isolated and inbred for many years.

Biologically significant features of this region of the Bam LINE are the PolII-like promoter signals and an open reading frame. Although our genomic consensus sequence features both the PolII-like signals and a significant open reading frame, we feel that the arguments associated with either feature are largely speculative in the absence of data directly implicating function. It is worthy of note, however, that should this PolII-like signal often promote an artifactual transcript that terminates at the reported polyadenylation site (15) approximately 1.3kb downstream (5), this may be a source of the concerted evolution by transposition, which is currently a popular theme. If this mechanism is operative, a higher copy number of the region represented by the transcript is predicted. However, experimental confirmation of this prediction is subject to numerous ambiguities.

We conclude from our results that the most significant aspect of this region of the repeated DNA sequence is its homogeneity with respect to length, evidenced by a dominant prototypic sequence. Apparently, the forces acting in the evolution of this repeated DNA sequence are conserving length and not necessarily bases, and we consider, as previously stated (5), that

this kind of conservation suggests a structural function for the Bam LINE of rodent.

ACKNOWLEDGEMENT

This work was supported in part by grant RR-5511 from the National Institutes of Health. We wish to thank Dr. Donald Robberson for his encouragement.

Abbreviations: bp = base pairs, kbp = kilo-base pairs, T_m = temperature midpoint of duplex strand separation.

REFERENCES

- 1) Singer, M., Thayer, R., Grimaldi, G., Lerman, M. and Fanning, T. (1983) Nucleic Acids Res. 11, 5739-5745.
- 2) Brown, S. and Piechaczyk, M. (1983) J. Mol. Biol. 165, 249-256.
- 3) Rogers, J. (1983) Nature 306, 113-114.
- 4) Meunier-Rotival, M. and Bernardi, G. (1984) Nucleic Acids Res. 12, 1593-1608.
- 5) Vizard, D. and Yarsa, J. (1984) Biochim. et Biophys. Acta 781, 143-152.
- 6) Brown, S. (1983) Gene 23, 95-97.
- 7) Martin, S., Voliva, C., Burton, F., Edgell, M. and Hutchison, C. (1984) Proc. Natl. Acad. Sci. U.S.A. 81, 2308-2312.
- 8) Maxam, A. and Gilbert, W. (1980) Methods in Enzymology 65, ed. Moldave, K., Nucl. Acids Pt. I., 499-560.
- 9) Rubin, C. and Schmid, C. (1980) Nucl. Acids Res. 8, 4613-4619.
- 10) Geck, P. and Nasz, I. (1983) Analytical Biochem. 135, 264-268.
- 11) Lund, T., Grosveld, F. and Flavell, R. (1982) Proc. Natl. Acad. Sci. U.S.A. 79, 520-524.
- 12) Klysik, J., Stirdivant, S. and Wells, R. (1982) J. Biol. Chem. 257, 10152-10158.
- 13) Drake, J., Glickman, B. and Ripley, L. (1983) American Scientist 71, 621-630.
- 14) Smith, T., Waterman, M. and Sadler, J. (1983) Nucl. Acids Res. 11, 2205-2220.
- 15) Fanning, T. (1983) Nucleic Acids Res. 11, 5073-5091.