Volume 11 Number 15 1983 Nucleic Acids Research

Size and structure of the highly repetitive BAM HI element in mice

Thomas G.Fanning

Department of Pathology, School of Medicine, University of California, Davis, CA 95616, USA

Received 6 June 1983; Accepted 7 July 1983

ABSTRACT

The BAM HI family of long interspersed DNAs in mice represent as much as 0.5% of the mouse genome. Cloned mouse DNA fragments which contain BAM HI/non-BAM HI junction sequences have been analyzed by restriction mapping and DNA sequencing. It has been found that BAM HI elements: (i) are approximately 7 kilobase pairs in size, (ii) are not bracketed by long repeated sequences analogous to the terminal repeats of proviruses and (iii) contain a poly-dA track at one end. The data strongly suggest that BAM HI elements arose by a process involving RNA intermediates. The beginning of the element, opposite the poly-dA track, contains a 22 base pair sequence exhibiting 65% homology to a ubiquitous mammalian sequence which may play a role in DNA replication (1). The poly-dA end of the element contains BAM5 and R sequences, both of which have been described previously (2,3).

INTRODUCTION

The fact that many organisms contain highly repetitive DNAs was established about fifteen years ago on the basis of DNA renaturation kinetics (4). Extensive analyses of these sequences in subsequent years has demonstrated that repetitive DNAs can be broadly grouped into several categories: multigene families which contain anywhere from several members to hundreds of members (5), satellite DNAs which usually consist of hundreds of short, tandemly linked sequences (6), short interspersed repetitive elements that are normally several hundred base pairs (bp) in size (7) and long interspersed repeats which appear to be thousands of base pairs (kbp) in size (8). Many short and long interspersed repetitive DNAs are present in as many as 10^4-10^6 copies per diploid cell.

The long interspersed repeats of mammals, designated as LINEs (9), have been extensively studied in rodents (2,10,11)

and primates (12,13). The most widely studied mouse LINE has been variously designated as the MIF-1 family (14) or the BAM HI family (11) and has been estimated to be approximately 3 kbp (14), or over 6 kbp (2,11), in size. In this report the term BAM HI will be used.

The BAM HI family of DNA sequences is highly polymorphic (10,11) and is present in as many as 10^4 copies per haploid mouse genome (10,11). A number of groups have analyzed portions of the BAM HI element (2,10,11,14,15,16,17) and several groups have demonstrated transcription of BAM HI sequences in mouse cells (2,11). As yet, however, no functional role has been demonstrated for this highly repetitive class of DNAs.

The experiments presented in this report were designed to examine certain structural aspects of BAM HI family members. It has been found that many BAM HI elements are large (ca. 7 kbp), homogeneous (80-95% sequence homology), unique in sequence (no repeat elements, tandem or otherwise) and often neighbor other, unrelated repetitive elements. The structural features exhibited by the BAM HI family of repetitive elements strongly suggests that they originally arose by, and perhaps are still propagated by, reverse transcription of RNA molecules, with subsequent insertion of DNA copies into the mouse genome. Such elements are referred to as "processed pseudogenes" or "processed genes" (18,19) and have been described in a number of systems.

MATERIALS AND METHODS

6.2 kbp Bgl I fragments were isolated from Bgl I-digested mouse genomic DNA as reported previously (2). The fragments were cleaved with Xba I plus Pst I and cloned into a Xba I/Pst I-cleaved pUC13 (20). One clone, pMXB6, contained an entire Xba I-Bgl I fragment (Fig. 1). Since the plasmid was not cleavable with Pst I, it appeared as though the Bgl I end of the fragment had ligated into the Pst I site of the plasmid.

Plasmids harboring mouse inserts containing BAM HI/non-BAM HI junction fragments were obtained by screening genomic mouse clones with either pMXB6 or pMBA14 insert DNAs (Fig. 1).

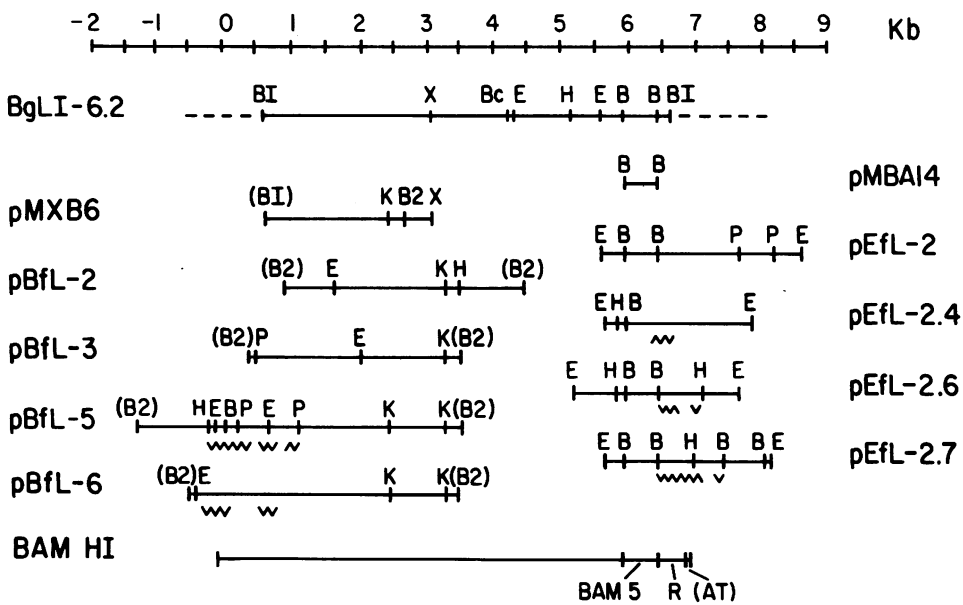

Fig. 1: Restriction maps of the mouse BAM HI element and BAM HI subclones. The upper map is the 6.2 kbp Bgl I fragment isolated from Bgl I-cleaved genomic mouse DNA. Restriction sites were determined by digestion of the isolated fragments, Southern blotting and probing with either labeled 6.2 kbp Bgl I DNA (see Fig. 2), pMXB6 insert DNA or pMBA14 insert DNA. The locations of the pMXB6 and pMBA14 subclones are shown directly below the 6.2 kbp Bgl I map. Plasmids containing mouse DNA were screened with pMXB6 insert DNA and four clones which contained homologous sequences were isolated and designated pBfl-2, 3, 5 and 6 (left side of figure). Similarly, four clones containing sequences homologous to the insert in pMBA14 were isolated and designated pEfl-2, 2.4, 2.6 and 2.7 (right side of figure). The very bottom line of the figure depicts the BAM HI element while the top line is the size scale (in kbp) and is zeroed on the beginning of the BAM HI element. Wavey lines show areas that have been sequenced. Restriction sites are: B, Bam HI; BI, Bgl I; BII, Bgl II; Bc, Bcl I; E, Eco RI; H, Hind III; K, Kpn I; P, Pst I; X, Xba I; (BI) a hybrid Bgl I/Pst I site; (B2) hybrid Bgl II/Bam HI sites.

pMXB6 and pMBA14 insert DNAs were excised from their respective plasmids with restriction enzymes, isolated by polyacrylamide gel electrophoresis, and labeled by nick translation (21). Mouse liver DNA (GR mouse strain) was digested to completion with either Eco RI or Bgl II and fractionated by agarose gel electrophoresis. Size classes of 2.5-3 kbp (Eco RI fragments)

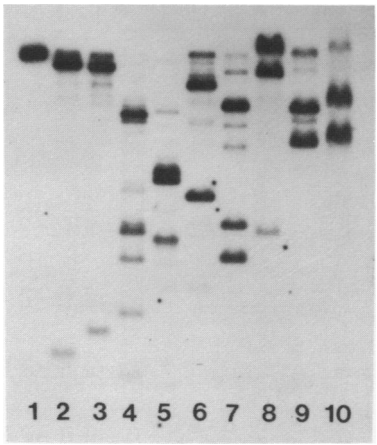

Fig. 2: Restriction profiles of the 6.2 kbp Bgl I fragment isolated from genomic mouse DNA. The fragment was digested with restriction enzymes, Southern blotted and probed with labeled 6.2 kbp Bgl I fragments. 1, no enzyme; 2, Bam HI; 3, Msp I; 4, Taq I; 5, Hae III; 6, Hind III; 7, Eco RI; 8, Pst I; 9, Xba I; 10, Bcl I. The locations of several of these sites are shown in figure 1.

or 4-5 kbp (Bgl II fragments) were cut from the gel and the DNA eluted. The Eco RI fragments were cloned in Eco RI-cleaved pUC9 while the Bgl II fragments were cloned in Bam HI-cleaved pUC13. Plasmids containing Eco RI inserts were screened (22) with labeled pMBA14 insert DNA. In a similar fashion, plasmids containing the Bgl II inserts were screened with the pMXB6 probe. Southern blotting and sequencing were done as described previously (2).

All experiments were performed under guidelines set forth by the NIH and the IBC-UC Davis.

RESULTS

Isolation and mapping the genomic 6.2 kbp Bgl I fragment.

Mouse liver DNA was cleaved with Bgl I and the fragments separated by agarose gel electrophoresis. An ethidium bromide-staining band which migrated to a position of approximately 6.2 kbp was excised from the gel and the DNA isolated. This family of DNA fragments has previously been shown to contain the highly repetitive BAM5 sequence (2).

Restriction mapping of the Bgl I fragments demonstrated their apparent homogeneity. Major fragments were seen following digestion with all of the enzymes that were tested (Fig. 2). The fragments were positioned on the map by Southern blotting and hybridization to cloned probes consisting of subfragments from either the right or left end of the Bgl I fragment (data

not shown).

Several enzyme digests gave rise to faint bands in addition to the major fragments (Fig. 2, lanes 4-7,9). Also, cloned fragments from different BAM HI family members exhibited numerous restriction site polymorphisms and variants (see below). It therefore seems safe to conclude that the Bgl I fragments, while fairly homogeneous in sequence, are not all identical. Similar conclusions have been reached by others (10,11).

The sizes of the restriction fragments, as well as Southern blotting data (not shown), leave little doubt that the 6.2 kbp Bgl I fragment contains subfragments described by numerous groups (10,11,14,15,16,17) and designated as either the MIF-1 family (14) or the BAM HI family (11).

Cloning and mapping BAM HI/non-BAM HI junction fragments.

The Bgl I sites which delinate the ends of the 6.2 kbp Bgl I fragment occur within the BAM HI element. In order to determine the size of the entire element, mouse DNA fragments containing BAM HI/non-BAM HI junction sequences were cloned and analyzed.

Mouse liver DNA was digested with either Eco RI or Bgl II, fractionated on agarose gels, and fragments of 2.5-3 kbp (Eco RI) or 4-5 kbp (Bgl II) were cloned in pUC plasmids. The Eco RI inserts were probed with labeled insert from the pMBA14 plasmid (BAM5 fragment (2)) while the Bgl II inserts were probed with a labeled Bgl I/Kpn I fragment isolated from the plasmid pMXB6 (see Fig. 1). Since the two probes originate from opposite ends of the 6.2 kbp Bgl I fragment and have no sequences in common (see below), they should detect inserts containing BAM HI/non-BAM HI junction sequences from either end of the BAM HI element.

Four clones that hybridized well with the pMXB6 probe were selected for study and designated as pBfl plasmids (Fig. 1). Likewise, four clones that hybridized well with the pMBA14 probe were isolated and designated as pEfl plasmids (Fig. 1). Two of the pBfl plasmids contained BAM HI/non-BAM HI sequences while two consisted only of BAM HI sequences. All four pEfl plasmids contained BAM HI/non-BAM HI sequences.

Establishing the boundries of the BAM HI element. Because of the structural similarities between BAM HI elements and processed pseudogenes this report will use the following conventions: all sequences at the left end of the BAM HI element well be referred to as 5'-sequences, while sequences at the right end of the element will be designated as 3'-sequences (see Fig. 1).

To determine the boundries of the BAM HI element three approaches were taken: (i) insert DNAs were digested with restriction enzymes, Southern blotted and probed with labeled total mouse DNA, (ii) subfragments from insert DNAs were labeled and used to probe Southern blots of genomic mouse DNA, and (iii) selected areas of the inserts were sequenced. In most cases only one or two of these approaches were used to characterize any given insert.

The 5'-end (left end) of the BAM HI element.

(a) Restriction mapping the pBfl inserts. The pBfl plasmids contained inserts which hybridized to the insert present in pMXB6 (Fig. 1). Since the inserts in the pBfl plasmids were present in the Bam HI site of pUC13, they were bracketed by the pUC13 cloning sites: Eco RI on one side and Pst I and Hind III on the other (20). Taking advantage of the strategic placement of these enzyme sites adjacent to the insert, a combination of single and double digests made the positioning of Eco RI, Bam HI, Hind III, Pst I and Kpn I sites within the inserts relatively straight-foreward (not shown).

To orient the inserts, pBfl plasmids were digested with various restriction enzymes, Southern blotted and probed with the 1.9 kbp Bgl I/Kpn I fragment (map positions in Fig. 1 (m.ps.) 0.6 to 2.5) from pMXB6, or with the 0.8 kbp Kpn I fragment (m.ps. 2.5 to 3.3) from pBfl-5. The resulting restriction maps are shown in Figure 1 and demonstrated that of the four cloned inserts, only those in pBfl-5 and pBfl-6 appeared to be of sufficient size to warrant further analysis.

(b) Repetitive sequences in pBfl-5. The plasmid pBfl-5 was digested with various enzymes, Southern blotted and probed with labeled total mouse DNA. This type of analysis makes use of the fact that only moderately to highly repetitive DNA sequences

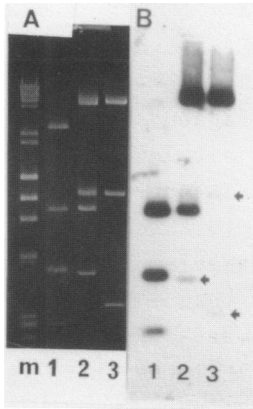

Fig. 3: Repetitive sequences in pBfl-5. pBfl-5 was cleaved with Hind III plus Pst I (lane 2) or with Hind III plus Bam HI (lane 3). As a hybridization control, and to provide size markers, pEfl-2.7 was cleaved with Eco RI plus Bam HI (lane 1). The digests were run on agarose gels, Southern blotted and probed with labeled total mouse DNA. Panel A shows the stained gel and panel B the autoradiogram. The sizes of the pEfl-2.7 fragments are (from top to bottom in lane 1) 900 bp, 507 bp and 274 bp. The arrows in lanes 2 and 3 are explained in the text. Lane m contains lambda-Hind III and ØX174-Hae III size markers.

will be present in sufficient concentration, and contain sufficient label, to hybridize and identify cloned restriction fragments (23). The results demonstrated that the 900 bp Pst I fragment (m.ps. 0.4 to 1.3) of pBfl-5 contained sequences that were highly repeated in the mouse genome and, thus, this fragment is undoubtedly situated within the BAM HI element (Fig. 3, lane 2).

The 500 bp Hind III/Pst I fragment (m.ps. -0.1 to 0.4) also hybridized a significant amount of labeled mouse DNA (Fig. 3, lane 2, arrow), but far less than the 900 bp Pst I fragment. The simplest explanation for this result is that the 500 bp fragment consists partly of BAM HI sequences and partly of non-BAM HI sequences, the latter being present as single-copy (or low-copy) sequences in the mouse genome.

Digestion of pBfl-5 with Hind III plus Bam HI gave rise to a 350 bp fragment (m.ps. -0.1 to 0.25) (Fig. 3, lane 3, lower arrow) which hybridized even less probe than the 500 bp Hind III/Pst I fragment. Since approximately 150 bp of the Hind III/Pst I fragment had been removed to produce the 350 bp Hind III/Bam HI fragment, and since the amount of hybridization to the smaller fragment had decreased by a factor of two or more (compare the intensities of the 500 bp and 350 bp fragments in Fig. 3, lanes 2 and 3), it was concluded that approximately 150 bp of repetitive sequences were still present on the 350 bp fragment. This suggested that the BAM HI element started somewhere between the Eco RI (m.p. 0) and Bam HI (m.p. 0.25)

sites on the pBfl-5 insert (see Fig. 1).

Weak hybridization of labeled total mouse DNA was detected to the 1.1 kbp Bgl II/Hind III fragment (m.ps. -1.2 to -0.1) from pBfl-5 (Fig. 3, lane 3, upper arrow: this fragment is released from the insert by virtue of a Hind III site on pUC13 which abuts the Bgl II/Bam HI hybrid site). One interpretation of this result is that BAM HI sequences extend into this area of the insert. This could be the case if, for example, a small portion of the BAM HI elements were longer than 7 kbp, or if a great deal of sequence heterogeneity between different BAM HI elements existed in this area, or both. While neither of these possibilities has been excluded, it seems more probable that sequences corresponding to a non-BAM HI repetitive element are present in the 1.1 Bgl II/Hind III fragment. Such non-BAM HI repetitive sequences have been found near the 3-'ends of some BAM HI family members and are within 500-1000 bp of the BAM HI element (see below).

(c) Sequence of the 5'-end of the BAM HI element. Partial sequence data was obtained from the Hind III site (m.p. -0.1) to the Pst I site (m.p. 1.3) of the pBfl-5 insert (not shown in toto, but see reference 24). In order to determine the exact 5'-end of the BAM HI element a 1.3 kbp Sau3A I fragment was isolated from pBfl-6. This fragment begins just to the right of the Eco RI site (m.p. -0.4) in pBfl-6 and extends 1.3 kbp farther to the right (m.p. 0.9). Approximately 250 bp from each end of the fragment were sequenced by the M13/dideoxy method and compared to the sequenced portion of pBfl-5. The Eco RI proximal portion of the 1.3 kbp Sau3A I fragment exhibited no sequence homology with pBfl-5. The Eco RI distal portion, however, showed a 93% homology with pBfl-5 sequences (m.ps. 0.65 to 0.9) (not shown (24)).

The 1.3 kbp Sau3A I fragment was restriction mapped with Msp I (Hpa II) to give the following map: Sau3A I - 194 bp - Msp I - 72 bp - Msp I - 142 bp - Msp I - ca. 900 bp - Sau3A I. Sequencing demonstrated that the 194 bp and 72 bp fragments had no homology to pBfl-5 sequences (not shown (24)). The sequence of the 142 bp Msp I fragment of pBfl-6 is shown in Figure 4A, together with the relevant sequence from pBfl-5.

```
(A)                                Eco RI
            -30        -20        -10         1         10
pBfl-5:  GCGGGAGACATCTTCGTTCCTGAATTCCACCTAGACTAGTCTGCACAGG
pBlf-6:  C    --  -  --  --  -   C -T GG- A T       A
              20         30         40         50         60
pBfl-5:  TGA--GTGTGGACTACAGAAGCTAACACCTTCTGGGACAGGCAGGAGCC
pBfl-6:     GA G   C  C      C  C   A          -        -T
               70         80         90         100
pBfl-5:  ACAGAGCTGCTGAGACAGT-CTCTGTTTCCGGCTCCAGACATGCGGG
pBfl-6:         C      G-  CA C  -    G   CG      -  C

(B)                 . 7 bp . .   14 bp    .
                                 A          A
          consensus:  CTACT-TG-GGAGGCTGAGGCGG
             pBfl-5:     G C  C C   -    TGT
             pBfl-6:  T  G C  A C   -    AG
```

Fig. 4: Sequences of the 5'-junction (left end) of the BAM HI element. (A) Comparison of sequences from pBfl-5 and pBfl-6 near the pBfl-5 Eco RI site (see Fig. 1). Nucleotide #1 is the beginning of the BAM HI element. (B) Comparison of the initial 22 bp of the pBfl-5 and pBfl-6 BAM HI elements with the 21 bp consensus sequence described by Jelinek et al. (1). The 21 bp is composed of two sections: a highly conserved 14 bp sequence and a less conserved 7 bp sequence. Positions #1 and #13 of the 14 bp sequence can be either A or G (1). Only sequence differences are shown in the figure and - indicates a missing nucleotide.

The 35 bp of pBfl-5 preceeding the BAM HI element (positions -35 to 0 in Fig. 4A) show little homology with the corresponding 25 bp of pBfl-6. There is a maximum of 55% homology between the two sequences prior to the beginning of the BAM HI element, however, this value is achieved only when numerous single and double base pair deletions are allowed in the pBfl-6 sequence (Fig. 4A). After position #1 the homology rises to 77% with very few insertions/deletions.

The BAM HI element appears to begin about 10 bp to the right of the Eco RI site of pBfl-5 (Fig. 4A). This is in excellent agreement with the data presented in Figure 3, which suggested that the 5'-end of the BAM HI element was situated somewhere near the middle of the 350 bp Hind III/Bam HI fragment (m.ps. -0.1 to 0.25) of the pBfl-5 insert.

The 5'-end of the BAM HI element contains sequences that are 65% homologous to sequences found in other repetitive and nonrepetitive DNAs (1) (Fig. 4B). These sequences may play

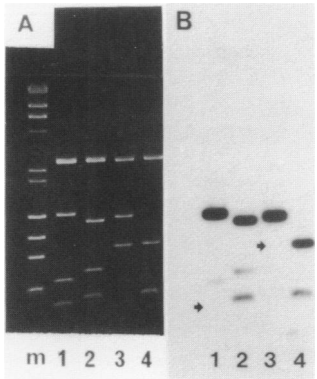

Fig. 5: Repetitive sequences in pEf1-2.6 and pEf1-2.7. pEf1-2.6 was cleaved with Eco RI plus Hind III (lane 1) or with Eco RI plus Bam HI (lane 2). pEf1-2.7 was cleaved with Eco RI plus Hind III (lane 3) or with Eco RI plus Bam HI (lane 4). The digests were run on agarose gels, Southern blotted and probed with labeled total mouse DNA. Panel A shows the stained gel and panel B the autoradiogram. Arrows are explained in the text. Lane m contains lambda-Hind III and ØX174-Hae III size markers.

some, as yet undetermined, role in DNA replication (1) (see DISCUSSION).

The 3'-end (right end) of the BAM HI element.

(a) Restriction mapping the pEf1 inserts. The pEf1 plasmids contained inserts which hybridized to the insert present in pMBA14 (Fig. 1). These inserts are located in the Eco RI site of pUC9 and, thus, are bordered by the Hind III, Bam HI and Pst I sites present in the cloning region of the plasmid (25). Making use of these sites, single and double digests of the plasmids with several enzymes established the restriction maps shown in Figure 1 (data not shown). An additional factor in establishing the pEf1 restriction maps was the paucity of restriction site polymorphisms when compared with the pBf1 plasmids. Three of the four inserts released the 507 bp BAM5 fragment (2) upon digestion with Bam HI. The single exception, pEf1-2.4, lacked a Bam HI site (m.p. 6.5) due to a single C to T transition within the restriction site (Fig. 7B).

(b) Repetitive sequences in pEf1 inserts. pEf1-2.6 and pEf1-2.7 were digested with restriction enzymes, Southern blotted and probed with labeled total mouse DNA. Both the 450 bp Hind III/Eco RI fragment (m.ps. 7.1 to 7.6; Fig. 5B, lane 1, arrow) of pEf1-2.6 and the 1.0 kbp Hind III/Eco RI fragment (m.ps. 7.0 to 8.0; Fig. 5B, lane 3, arrow) of pEf1-2.7 were not identified by the probe. Therefore, these two fragments must be present in low copy number in the mouse genome and not part of the BAM HI element.

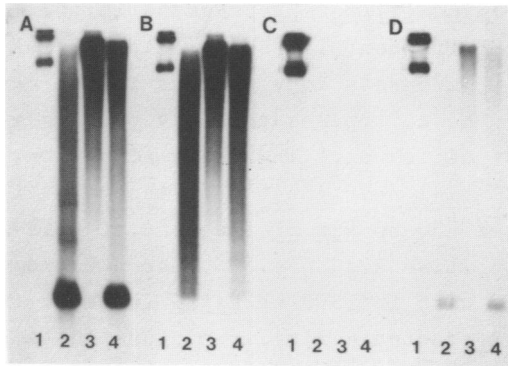

Fig. 6: Southern blots of genomic mouse DNA probed with fragments from pEf1-2.7. Approximately 8 ug of mouse liver DNA was digested with Bam HI plus Hind III (lane 2), Hind III (lane 3) or Bam HI (lane 4), Southern blotted and probed with labeled fragments from pEf1-2.7. As a control for blotting and hybridization approximately 500 ng of pEf1-2.7 plasmid (consisting of both monomeric and dimeric forms) was run in lane 1. The blots were probed with: (A) a 507 bp Bam HI fragment (m.ps. 6.0 to 6.5; the BAM5 sequence (2)), (B) a 510 bp Bam HI/Hind III fragment (m.ps. 6.5 to 7.0; an R element (3)), (C) a 450 bp Hind III/Bam HI fragment (m.ps. 7.0 to 7.45) and (D) a 550 bp Bam HI fragment (m.ps. 7.45 to 8.0).

(c) Hybridization of pEf1 subfragments with mouse DNA. BAM HI elements are highly repetitive interspersed DNAs in the mouse genome. Probing Southern blots of genomic mouse DNA with subfragments from the pEf1 inserts should therefore identify which fragments are homologous to repetitive DNAs and which are not. A series of such Southern blots is shown in Figure 6.

pEf1-2.7, as well as subfragments of the pEf1-2.7 insert, were digested with Bam HI and Hind III to give four fragments ranging from 450-550 bp in size. Each fragment was isolated, labeled by nick translation and used to probe Southern blots of genomic mouse DNA digested with the same enzymes. Two of the fragments, the 507 bp Bam HI fragment (m.ps. 6.0 to 6.5) and the 510 bp Bam HI/Hind III fragment (m.ps. 6.5 to 7.0) were highly repetitive and contained BAM HI sequences (Fig. 6A, 6B) The 507 bp Bam HI fragment is the BAM5 sequence reported previously (2), while the 510 bp Bam HI/Hind III fragment is an R element (3)(see below).

The 450 bp Hind III/Bam HI fragment (m.ps. 7.0 to 7.45) and 550 bp Bam HI fragment (m.ps. 7.45 to 8.0) of pEfl-2.7 are present in low to moderate copy numbers in the mouse genome (Fig. 6C, 6D). The 450 bp sequence may be single-copy, but the 550 bp sequence is not. The bands at 550 bp, plus the hybridizable material near the top of the gel (Fig. 6D), indicate that the 500 bp Bam HI fragment of pEfl-2.7 was a member of a repetitive sequence. Southern blots of genomic mouse DNA probed with the 500 bp Pst I fragment (m.ps. 7.7 to 8.2) of pEfl-2 presented a similar picture (data not shown). Since the 550 bp Bam HI and 500 bp Pst I fragments failed to hybridize with one another or with any other pEfl or pBfl sequences (not shown) it is unlikely that these repetitive DNAs are members of the BAM HI family.

The non-BAM HI portions of pEfl-2 and pEfl-2.7 contained sequences that are repetitive in the mouse genome. pBfl-5 also appeared to contain sequences, outside of the BAM HI element, which are repetitive (see Fig. 3). Although this limited number of examples does not make a strong case, the data do suggest that many BAM HI elements are bordered by other, unrelated repetitive elements. Whether BAM HI elements and non-BAM HI repetitive elements are normally separated by short stretches of low-copy DNA (as is the case with pEfl-2.7, see Fig. 6) is unknown.

(d) Sequence of the 3'-end of the BAM HI element. Sequence data was obtained from pEfl-2.4, pEfl-2.6 and pEfl-2.7 as indicated in Figure 1 (wavey lines). Approximately 150 bp of pEfl-2.4 were sequenced, starting at a Hae III site (m.p. 6.6) and ending within the BAM5 sequence (m.p. 6.4). pEfl-2.6 has been sequenced (with several minor gaps) from the Bam HI site (m.p. 6.5) to the Hind III site (m.p. 7.1). pEfl-2.7 was sequenced (with several minor gaps) from the Bam HI site (m.p. 6.5) to the Bam HI site (m.p. 7.4). Relevant portions of the sequences are shown in Figure 7.

Figure 7A shows the end of the BAM HI homology region in pEfl-2.6 and pEfl-2.7. Both members of the BAM HI family end in d(AT)-rich regions of DNA. pEfl-2.6 contains eight dA, a dG and then three dA residues (positions 6989 to 7000).

```
(A)            6960       6970       6980       6990       7000
pEf1-2.6:  TAGCATTGGAAATGTAAATGAGGAAAATACCTAATTAAAAAAAAGAAA
pEf1-2.7:       G    G                          A    T TTTTT
                 7010       7020       7030
pEf1-2.6:  TATTTAGATATATCTTAACTACCTAATTTGTG
pEf1-2.7:  A AAGTTT CAGGA   GT   TAGTGC  G

(B)              10         20         30         40         50
 R cons.:  GGATCCATCCCATAATCAGCCACCAAACCCANACACTATTGCATATNCNA
pEf1-2.7:                      TT     TG T     C       C C
pEf1-2.6:                       T      G T     C         C
pEf1-2.4:       T     - T  T   TT      GAT     C    T    C
                 60         70         80         90         100
 R cons.:  ANAAGATTTTGCTGAAAGGANNCTGATATAGCTGTCTCTTGTGAGACTAT
pEf1-2.7:  G   -      T  T        A         C     A
pEf1-2.6:  G       G      T                A      AC
pEf1-2.4:  G    CG    T  C       GA                          G

                        -----  291 bp  -----

                 401        411        421        431    _   441
 R cons.:  AAAAGTGGGAGTGGGTGGGTAGGGGAGTNGGGAGGGGAGGGTATGGGGGA
pEf1-2.7:              C     T     T       ---T
pEf1-2.6:                          GA     --- A
                 451        461        471        481        491
 R cons.:  CTTTTGGGATAGCATTTGAAATGTAAATGAGGAAAATATCTAATAAAAAA
pEf1-2.7:                G G  G                  C          T
pEf1-2.6:                  G                     C     T
```

Fig. 7: Sequences at the 3'-junction (right end) of the BAM HI element. (A) Comparison of sequences from pEf1-2.6 and pEf1-2.7 before and after the poly-dA tracks. The end of the BAM HI element has been numbered 7000 but this should not be taken as definitive since the exact length of the element is unknown. (B) Comparison of the consensus R sequence (3) with sequences in pEf1-2.4, pEf1-2.6 and pEf1-2.7. The numbering conforms to that of the R consensus sequence (3). The Bam HI site (nucleotides 1-6) is located at map position 6.5 in figure 1. Only sequence differences are shown in the figure and - indicates a missing nucleotide.

pEf1-2.7 contains two dA, a dT and then five dA residues (positions 6985 to 6992). Poly-dA tracks of variable length have been described at the borders of other genetic elements (7,18,19) and may result from a mode of DNA amplification involving an RNA intermediate (see DISCUSSION).

Sequences lying between the BAM5 fragment (m.ps. 6.0 to 6.5) and the junction of BAM HI DNA and non-BAM HI DNA are presented (in part) in Figure 7B and compared with the consensus

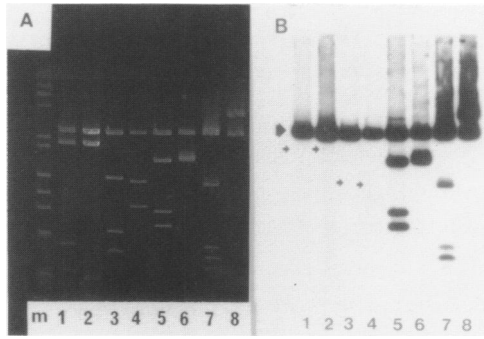

Fig. 8: Absence of shared sequences between pBlf and pEfl plasmids. Equal amounts of the four pBfl plasmids were mixed and labeled by nick translation. This probe was used on a Southern blot containing restriction enzyme-digested pEfl and pBfl plasmids. Panel A shows the stained gel and panel B the autoradiogram. Lanes 1-4 are pEfl-2, 2.4, 2.6 and 2.7 respectively. Lanes 5-8 are pBfl-2, 3, 5 and 6 respectively. All plasmids were cleaved with Eco RI plus the following: Pst I (pEfl-2, PBfl-3, 5 and 6), Hind III (pEfl-2.6, 2.7 and pBfl-2), Bam HI (pEfl-2.4). Small arrows in lanes 1-4 designate insert fragments containing BAM HI/non-BAM HI junction sequences. The large arrow shows the pUC plasmid band.

sequence of the mouse R element (3). The final 490 bp of the BAM HI family of repetitive DNAs are 80-90% homologous to the R element consensus sequence. This result extends the finding of Wilson and Storb (26) who demonstrated that several BAM5-R composite elements were present near mouse immunoglobulin light chain genes. Since Wilson and Storb's data suggested that these BAM5-R composite elements were not components of larger repetitive DNAs (ie. BAM HI elements) it appears as though sequences at the 3'-end of the BAM HI element can exist as separate entities in the mouse genome (see DISCUSSION).

Although only the initial and final 100 bp of R element homology are shown in Figure 7B, the entire R element present in pEfl-2.7 has been sequenced. The remaining 290 bp also show about 85% homology to the R element consensus sequence (not shown (24)).

The 3'-end and 5'-end of the BAM HI element do not share extensive sequence homology. A computer search failed to detect 5'-BAM HI sequences showing extensive homology to the R element

consensus sequence. However, to test the possibility of homology outside of the sequenced regions the four pBfl plasmids were mixed together, labeled by nick translation and used as probe on a Southern blot containing both pEfl and pBfl plasmids. No hybridization between the 5'-sequences present in the pBfl plasmids and the 3'-sequences present in the pEfl plasmids could be detected (Fig. 8).

This analysis does not exclude the possibility that short regions of good homology, or long regions of poorer homology, are shared between the 5'- and 3'-ends of the BAM HI element. Rather, the results presented in Figure 8 demonstrate that no long regions of good homology exist and that BAM HI elements are not bracketed by long terminal repeats of the type seen with proviruses and intracisternal A-particle genes (27,28,29).

DISCUSSION

Mouse BAM HI elements are interspersed repetitive DNAs present in approximately 10^4 copies per haploid mouse cell. The elements are about 7 kbp in size, contain no duplicated sequences and end in poly-dA tracks. These structural features are shared by "processed pseudogenes" and "processed genes". Processed pseudogenes and genes are thought to arise by the reverse transcription of cellular RNAs with subsequent insertion of DNA copies into the host genome (18,19,30). Processed immunoglobin, beta-tubulin, metallothionein, alpha-tubulin and alpha-globin genes have been described in humans, rats and mice (18). In addition, human snRNA pseudogenes (30) and the ubiquitous Alu (19,30) and mouse B1/B2 (31) sequences are thought to be, or have been, propagated by such a mechanism. Processed pseudogenes and genes share several characteristics in common, although not all of these characteristics are found in any given case (18): the genes lack introns, possess a poly-dA track at one boundary, and are bordered by short direct repeats of host DNA.

The presence of a poly-dA track at one end of the BAM HI elements is compatible with the idea that they multiplied by a mechanism similar to that suggested for processed pseudogenes. More significantly, both R elements (3) and BAM5-R composite

elements (26) possess poly-dA tracks and are bracketed by short direct repeats. This strongly suggests that subsegments of the BAM HI element propagate by means of RNA intermediates. Therefore, the present data is certainly compatible with the view that BAM HI elements multiplied throughout the mouse genome by a mechanism involving the reverse transcription of 7 kbp BAM HI RNAs. Such a process may still occur in mice since BAM HI transcripts are easily detected in mouse cells (2,11). Whether short direct repeats flank 7 kbp BAM HI DNAs is currently under investigation.

The reverse transcription of cellular messenger RNAs followed by insertion of DNA copies into the host genome appears to be a rare and capricious event (18). Yet, reverse transcription of genomic RNA followed by DNA insertion is part of the life cycle of retroviruses (28) and often results in the accumulation of dozens of proviral genomes in the host cell (32). Similarly, intracisternal A-particle genes may have propagated by means of RNA intermediates (33) resulting in hundreds of copies per cell (34). The BAM HI element (as well as Alu, B1, etc.) could be viewed as an extreme case of this mode of amplification which resulted in the accumulation of thousands of copies per cell. Since BAM HI-homologous sequences are present in high copy numbers in rats (2) these amplification events probably occured more than 10^7 years ago when mice and rats evolved into distinct species (35).

The data presented in Figure 3 demonstrates the highly repetitive nature of the 900 bp Pst I fragment (m.ps. 0.4 to 1.3) of pBfl-5 and suggests that many BAM HI elements are full length, ie. 7 kbp. However, the under-representation of this 900 bp Pst I fragment, when compared to the 900 bp Bam HI fragment (m.ps. 6.5 to 7.4) of pEfl-2.7 (see Fig. 3), implies that sequences present at the 5'-end of the BAM HI element are less reiterated in the mouse genome than sequences at the 3'-end of the element. Thus, a substantial proportion of BAM HI elements may be truncated at the 5'-end and, indeed, this view appears to be substantiated by the data of Gebhard et al. (3) and Wilson and Storb (26).

These two groups have characterized a number of R elements

(3) and BAM5-R composite elements (26) in mice. In most cases these R and BAM5-R elements appeared to be bracketed by single-copy (or low-copy) DNA sequences indicating that sequences derived from the 3'-end of BAM HI elements exist in the mouse genome as independent entities. If BAM HI elements, or sub-components of BAM HI elements, propagated through RNA intermediates this finding could be explained in several ways. Preferential losses of 5'-sequences during reverse transcription of 7 kbp BAM HI RNA would result in a greater proportion of 3'-DNA sequences being inserted into the mouse genome. Alternatively, many 7 kbp BAM HI elements might contain transcription initiation sites at points within the element. If this were the case then excessive transcription of 3'-sequences, relative to 5'-sequences, would be the result.

The models described above predict a higher copy number for the most 3'-proximal sequences of the BAM HI element, ie. the R sequences. This prediction is at odds with the results of Wilson and Storb (26) who found about $2x10^4$ BAM5 and 10^4 R sequences per haploid mouse genome based on dot blot hybridization. However, using a similar technique Gebhard et al. (3) arrived at a value of approximately 10^5 R sequences per haploid mouse genome. Differences in the hybridization conditions used by the two groups may be the cause of this descrepancy (26) and, indeed, some R elements are sufficiently divergent from the consensus sequence so as to preclude hybridization with a "standard" R element probe (3). The same situation is found with regard to BAM5 sequences. In a survey of cloned mouse DNA fragments, 13 of 17 clones which hybridized to a BAM5 probe (pMBA14; see Fig. 1) did so moderatly or poorly suggesting considerable divergence among BAM5 sequences (T.G. Fanning, unpublished observations). Thus, there may be many more copies of BAM5 and R sequences per haploid mouse cell than reported (26) although, because of sequence divergence, exact copy numbers may not be quantifiable by standard hybridization techniques. As suggested by Gebhard et al. (3) for R sequences, a value of 10^5 BAM5 sequences per haploid mouse cell is consistent with much of the data.

The beginning of the BAM HI element contains a 22 bp

sequence sharing 65% homology with a consensus sequence described by others and which may play a role in DNA replication (1). It is possible that the sequence homology seen in Figure 4B is simply the result of happenstance, however, the fact that it occurs directly at the beginning of the BAM HI element suggests otherwise. Although the original description of the 14 bp consensus sequence (Fig. 4B) emphasized its possible functional role in DNA replication (1), in the years since this report no definitive experimental results have emerged which support this contention (C. Schmid, personal communication). Thus, the functional role (if any) of this sequence, in BAM HI elements and in other repetitive and nonrepetitive DNAs, remains to be elucidated.

ACKNOWLEDGEMENTS

I thank G. Heidecker and R. D. Cardiff for suggestions and discussion. Supported by PHS grant 5 RO1 CA21454 from the NIH.

REFERENCES

1. Jelinek, W.R., Toomey, T.P., Leinwand, L., Duncan, C.H., Biro, P.A., Choudary, P.V., Weissman, S.M., Rubin, C.M., Houck, C.M., Deininger, P.L. and Schmid, C.W. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 1398-1402.
2. Fanning, T.G. (1982) Nuc. Acids Res. 10, 5003-5013.
3. Gebhard, W., Meitinger, T., Hochtl, J. and Zachau, H.G. (1982) J. Mol. Biol. 157, 453-471.
4. Britten, R.J. and Kohne, D.E. (1968) Science 161, 529-540.
5. Jeffreys, A.J. (1981) In: Genetic Engineering 2 (R. Williamson, ed.). Academic Press, New York. pp. 1-48.
6. Brutlag, D.L. (1980) Ann. Rev. Genet. 14, 121-144.
7. Jelinek, W.R. and Schmid, C.W. (1982) Ann. Rev. Biochem. 51, 813-844.
8. Singer, M.F. (1982) Int. Rev. Cytol. 76, 67-112.
9. Singer, M.F. (1982) Cell 28, 433-434.
10. Meunier-Rotival, M., Soriano, P, Cuny, G., Strauss, F. and Bernardi, G. (1982) Proc. Natl. Acad. Sci. U.S.A. 79, 355-359.
11. Soriano, P., Meunier-Rotival, M. and Bernardi, G. (1983) Proc. Natl. Acad. Sci. U.S.A. 80, 1816-1820.
12. Grimaldi, G. and Singer, M.F. (1983) Nuc. Acids Res. 11, 321-338.
13. Manuelidis, L. (1982) Nuc. Acids Res. 10, 3211-3219.
14. Brown, D.D.M. and Dover, G. (1981) J. Mol. Biol. 150, 441-466.
15. Cheng, S.M. and Shildkraut, C.L. (1980) Nuc. Acids Res. 8, 4075-4090.
16. Heller, R. and Arnheim, N. (1980) Nuc. Acids Res. 8, 5031-5042.

17. Manuelidis, L. (1980) Nuc. Acids Res. 8, 3247-3258.
18. Sharp, P.A. (1983) Nature 301, 471-472.
19. Jagadeeswaran, P., Forget, B.G. and Weissman, S.M. (1981) Cell 26, 141-142.
20. pUC13 is a derivative of pUC8 - see reference 25 - that contains a Xba I site in the plasmid cloning region. pUC13 was constructed and generously furnished by G. Heidecker.
21. Rigby, P.W., Dieckmann, M., Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113, 237-251.
22. Hanahan, D. and Messelson, M. (1980) Gene 10, 63-67.
23. Cochet, M., Gannon, F., Hen, R., Maroteaux, L., Perrin, F. and Chambon, P. (1979) Nature 282, 567-574.
24. Sequence data referred to, but not presented in this report, will be furnished upon written request.
25. Vieira, J. and Messing, J. (1982) Gene 19, 259-268.
26. Wilson, R. and Storb, U. (1983) Nuc. Acids Res. 6, 1803-1807.
27. Shimotohno, K., Mijutani, S. and Temin, H.M. (1980) Nature 285, 550-554.
28. Varmus, H.E. (1982) Science 216, 812-820.
29. Cole, M.D., Ono, M. and Huang, R.C.C. (1981) J. Virol. 38, 680-687.
30. Van Arsdell, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Manser, T. and Gesteland, R.F. (1981) Cell 26, 11-17.
31. Kramerov, D.A., Lekakh, I.V., Samarina, O.P. and Ryskov, A.P. (1982) Nuc. Acids Res. 10, 7477-7491.
32. Ringold, G.M., Shank, P.R. and Yamamoto, K.R. (1978) J. Virol. 26, 93-101.
33. Yang, S.S. and Wivel, N.A. (1973) J. Virol. 11, 287-298.
34. Lueders, K.K. and Kuff, E.L. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 3571-3575.
35. Rice, N.R. (1972) In: Evolution of Genetic Systems (H.H. Smith, ed.). Gordon and Breach, New York. pp. 44-79.