

Structure and Evolution of Two Related Transcription Units of Epstein-Barr Virus Carrying Small Tandem Repeats

GERHARD LAUX,* U. KARL FREESE,† AND GEORG W. BORNKAMM

Institut für Virologie, Zentrum für Hygiene, Universität Freiburg, D-7800 Freiburg, Federal Republic of Germany

Received 25 March 1985/Accepted 19 August 1985

Two regions of the Epstein-Barr virus (EBV) genome carrying partially homologous clusters of short tandem repeats (*NotI* and *PstI* repeats) flanked by 1044 and 1045 base pairs with almost complete homology (D_L and D_R , left and right duplication, respectively) were most abundantly transcribed into poly(A)⁺ mRNA after induction with the tumor promoter 12-*O*-tetradecanoyl-phorbol-13-acetate. The nucleotide sequence of both repeat clusters and the conserved upstream regulatory sequences from the M-ABA EBV strain are presented. Nearly the whole part of the sequences coding for the RNAs is covered by the *NotI* and *PstI* repeats, respectively. The regulatory sequences for these genes are located in the homologous regions of 1044 and 1045 base pairs (D_L and D_R , respectively). A CAAT box, a TATA box, and other herpes simplex virus-like elements were identified for both transcription units. The initiation points and the 3' ends of both inducible RNAs were mapped by S1 nuclease analysis. Both genes have open reading frames and may potentially code for proteins with repetitive amino acid compositions. The structure of these two inducible EBV genes is discussed, and an evolutionary model is proposed for the generation of gene duplication in the M-ABA strain of EBV.

The Epstein-Barr virus (EBV) is a lymphotropic human herpesvirus which is carried latently by most human adults. It is the etiological agent of infectious mononucleosis and has been implicated in two different human tumors: Burkitt's lymphoma and nasopharyngeal carcinoma (for a review, see reference 11). Additionally, the virus immortalizes human B lymphocytes in tissue culture.

The EBV genome from virus particles is a linear, double-stranded DNA molecule of about 175 kilobase pairs (kbp). It is characterized by a number of different repetitions. The termini consist of tandem repeats of 538 base pairs (bp). A variable number of large internal repeats of 3,072 bp joins a short and a long unique region. Several different other repeats are interspersed in the genome (2). Two clusters of small tandem repeats of 125 and 102 bp show partial homology and have the same orientation in the genome (8, 15, 17, 18, 20, 21). Each cluster is flanked by a highly conserved region of about 1 kbp. These left and right duplicated regions (D_L and D_R , respectively) are located about 100 kbp apart from each other in the viral genome (36). The structure of the viral genome is schematically shown in Fig. 1.

The repeat clusters of D_L and D_R belong to the most abundantly transcribed sequences of EBV. Transcription occurs only during the early and late phase of the lytic cycle of the virus. The RNA is not constitutively expressed but requires spontaneous or experimental induction (e.g., by the tumor promoter 12-*O*-tetradecanoyl-phorbol-13-acetate (TPA)). In both transcription units, transcription proceeds from right to left starting in the region of perfect homology, extending through the repeat clusters, and terminating at the left-hand side of the repeats. The RNA is poly(A)⁺ and can be purified from the polyribosomal fraction of the cytoplasm. The polymorphism in the length of the RNAs in different EBV strains reflects the polymorphism in the number of the repeats (15). The sequence analysis of the *NotI* and *PstI* repeats revealed open reading frames in both repeat clusters,

potentially giving rise to a protein with a repetitive amino acid structure (2, 15, 20).

We studied in detail the structure of the duplicated regions, including the *NotI* and *PstI* repeat clusters, the initiation sites of RNA synthesis, and the upstream flanking sequences, which are probably the targets for the regulated expression of these inducible genes. For an analysis of the evolutionary relationship it was important to study both regions within one virus isolate. In this report the complete nucleotide sequence of both regions from EBV strain M-ABA is presented. This is a transforming, nondefective virus strain which was originally derived from the tumor cells of a nasopharyngeal carcinoma (7) and has the structural organization of an EBV prototype (3, 34). The structure of the genes, including transcription regulatory elements and their potential coding capacity, is discussed, and a model for their evolution is proposed.

MATERIALS AND METHODS

DNA clones. Subclones containing D_L and D_R were constructed from a cosmid library of the M-ABA strain of EBV (34). For the analysis of D_R and the *PstI* repeats the *BglII* K fragment was subcloned into pHC79 and sequenced.

For analysis of the *NotI* repeats and D_L , the 4.5-kbp fragment overlapping between *BamHI*-H1 and *BglII*-C was cloned into pHC79. Subclones were constructed by digestion with *BamHI*-*BglII*-*SacI*, *BamHI*-*BglII*-*SacII*, and *BamHI*-*BglII*-*AvaI* followed by generating blunt ends with T4 DNA polymerase (Pharmacia Fine Chemicals) and cloning into the *HincII* site of pUC8 (40).

Plasmid DNA was prepared by the cleared lysate technique of Clewell and Helinski (5) followed by two consecutive cesium chloride-ethidium bromide gradients.

DNA sequence analysis. Both strands of the virus-specific DNA inserts were sequenced from appropriate restriction sites by the method of Maxam and Gilbert (31). Restriction fragments were labeled either at their 5' ends with T4 polynucleotide kinase (Bethesda Research Laboratories, Inc., Gaithersburg, Md.) and [γ -³²P]ATP (Amersham Corp., Arlington Heights, Ill.) after pretreatment with calf intestine alkaline phosphatase (Boehringer Mannheim) or at their 3'

* Corresponding author.

† Present address: Deutsches Krebsforschungszentrum, D-6900 Heidelberg, Federal Republic of Germany.

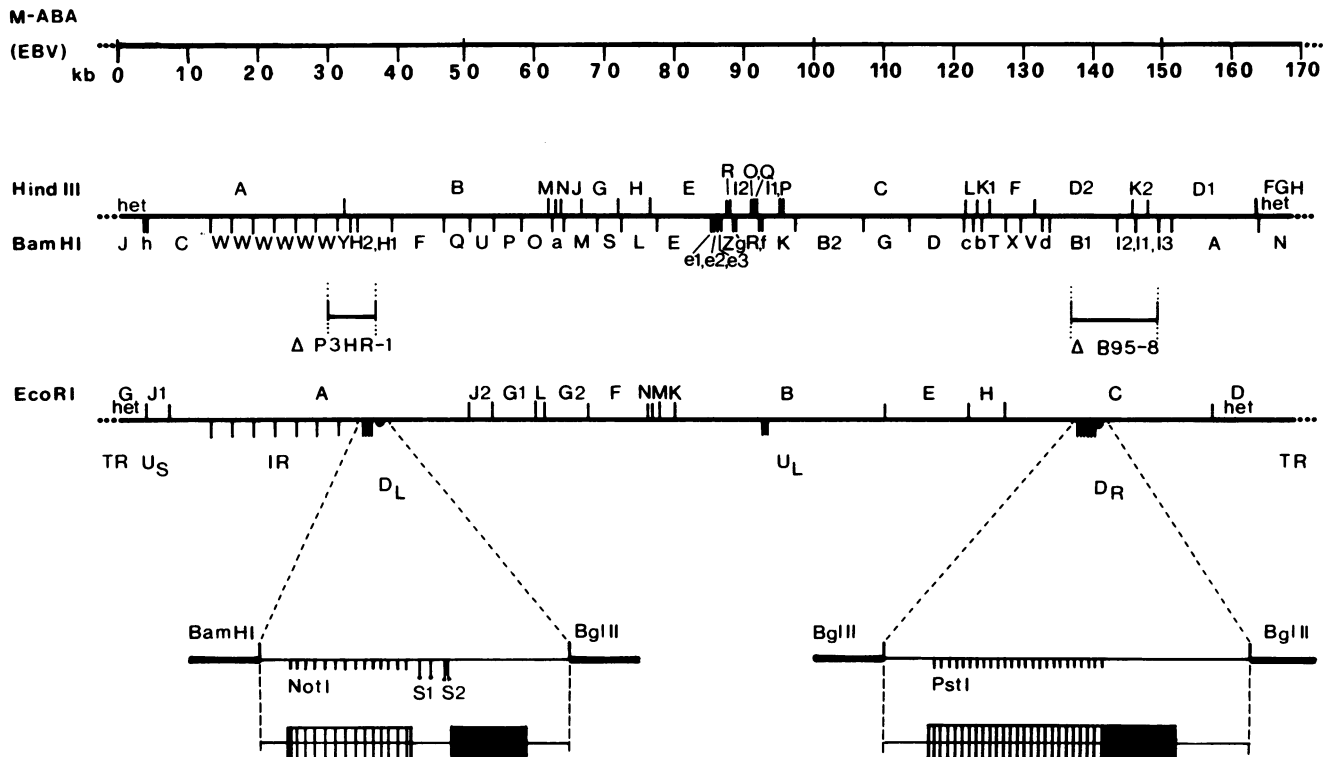


FIG. 1. Schematic diagram of the M-ABA EBV genome with restriction enzyme sites for *Hind*III, *Bam*HI, and *Eco*RI. The DNA consists of terminal repeats (TR) and a short (U_S) and a long (U_L) unique region joined by large internal repeats (IR). In the long unique region are two regions with sequence homology (D_L and D_R). Their organization in the M-ABA EBV clones pM800 (D_L) and pMB2-K (D_R) is shown in the lowest line. The duplications are indicated by black boxes, and the tandem repetitions are indicated by hatched boxes. The thick lines represent parts of the pHC79 vector. Relevant restriction enzyme sites are also shown (S1, *Sac*I; S2, *Sac*II). Deletions (Δ P3HR-1 and Δ B95-8) in the genome of P3HR-1 and B95-8 EBV strains are indicated.

ends by deoxyterminal transferase (Bethesda Research Laboratories) and [α -³²P]dideoxyATP (Amersham). Labeled DNAs were recut with appropriate restriction enzymes, and the fragments were isolated by agarose gel electrophoresis. If one of the two fragments was smaller than 20 bp, it was not removed from the sequencing reaction. Otherwise strands were separated by electrophoresis in 5% denaturing polyacrylamide gels.

Sequences of restriction fragments were assembled and analyzed by the aid of computer programs developed by Staden (38) and Kröger and Kröger-Block (25, 26).

S1 nuclease mapping. Two micrograms each of plasmid pM800 and pMB2-K DNAs were used for S1 nuclease analysis. For the mapping of the RNA initiation sites, plasmid DNAs digested with *Kpn*I were labeled by a replacement reaction with T4 DNA polymerase (30). The 3' exonuclease activity of the enzyme first excised about 400 nucleotides from each 3' end. The following polymerase activity incorporated the added labeled deoxyribonucleotide triphosphates.

The labeled DNA was separated from unincorporated nucleotides by Sephadex G-50 chromatography (Pharmacia Fine Chemicals). After they were recut with appropriate enzymes, the labeled probes were isolated by agarose gel electrophoresis.

S1 nuclease analysis was done as described by Favaloro et al. (14). Briefly, 1 μ g of each probe was hybridized with 100 μ g of cytoplasmic RNA isolated from TPA-induced M-ABA cells (15) in a total volume of 30 μ l of 80% deionized formamide for 15 h at 66°C. Digestion was carried out with

6,000 U of S1 nuclease (Boehringer Mannheim) in 400 μ l at 37°C for 75 min. The protected probe was analyzed on a 6% denaturing polyacrylamide gel followed by autoradiography.

For the mapping of the 3' ends of the RNAs, the plasmid DNA of pM800 was digested with *Bgl*II, and the plasmid DNA of pMB2-K was digested with either *Pst*I or *Sst*I.

Labeling, preparation, and S1 nuclease analysis were done as described above, except that the hybridization was carried out at 56, 60, and 52°C, respectively, taking into account the different G+C contents.

RESULTS AND DISCUSSION

The transcription unit containing the *Pst*I repeats. The strategy for sequencing of the *Bgl*II K fragment is shown in Fig. 2. The sequence is shown in Fig. 3. This fragment contains the *Pst*I repeat cluster first described by Hudewentz et al. (18). It consists of 25.6 repeat units of 102 bp each with an average G+C content of 84.3%. The sequence of the repeats shown in Fig. 3 was deduced from 4.2 individual repeat units by sequencing into the repeat cluster from both sides. In addition, both strands of the 102-bp repeat fragments generated by digestion with *Ava*I and *Pst*I, respectively, were sequenced. The difference of our sequence to that of Dambaugh and Kieff (8) derived from the AG876 EBV strain has been documented and discussed elsewhere (15). Even though the whole repeat cluster was not sequenced we have some evidence for single base pair variations among the individual repeats. *Apa*I cuts twice within the repeats: once in the outmost right repeat and once in the fourth repeat starting from the left-hand side. Both



FIG. 3. The nucleotide sequence of the *Bg/III* K fragment of M-ABA EBV DNA. The borders of the *PsfI* repeats and the homologous region (*D_R*) are indicated by brackets and vertical bars, respectively. Dashes represent parts of the *PsfI* repeat cluster that were not sequenced. The TATA-like, CAAT, and AATAAA sequences referred to in the text are boxed, and AC strings are underlined. The TAATGARAT elements are marked by hyphens. The 5' end of the message was determined by S1 nuclease analysis and is marked by a triangle, with a small arrow showing the direction of transcription. The long arrows in *D_R* illustrate two palindromic sequences.

boundary between the spacer and *D_L* whereas the *PsfI-KpnI* fragment spanned the boundary between the *PsfI* repeats and *D_R*. The size of the transcripts, as revealed by Northern blotting, suggests that both inducible RNAs are initiated in the conserved regions (15). By hybridization of cytoplasmic RNA with the *D_L* probes, fragments of 162 and 183 bases

were protected by the *SacII-KpnI* probe, and fragments of 402 and 423 bases were protected by the *SacI-KpnI* probe (Fig. 5, lanes 1 and 3). This indicates that RNA is initiated at two distinct sites in *D_L*, 21 bases apart from each other. Two additional bands of about 570 and 590 bases were visualized by the *SacI-KpnI* probe. They were generated by partial

A

```

TCACGGGGGAGGACCGCGCCGAGCCACCAGGGGCCCGGGGGTGGGGGGTGCCTCCAGGCCGGACCCCTGGTGCCAGGCAGGGACCCCTGCGCCACC
          GGGGGTGCCTCCAGGCCGGACCCCTGGTGCCAGGCAGGGACCCCGGCCACC
          3242
CGCTTCATGGGGGGGAGGCCGCGCAAGGACCGCGGGCCGGCTGGGAGGTGTGCACCCCGGAGCGCTGGGAGACBCGGGAGCCGGCCGGCTGGC
CGCTTCATGGGGGGGAGGCCGCGCAAGGACCGCGGGCCGGCTGGGAGGTGTGCACCCCGGAGCGCTGGGAGACBCGGGAGCCGGCCGGCTGGC
CTTTTATATCTCTTTTTGGGGTCTCTGGTAATACTTCAAGGTTTGCACAGGAGTGGGGGCTTCTATTGGTTAATTCAGGTTGTCTATTTAGCCCG
CTTTTATATCTCTTTTTGGGGTCTCTGGTAATACTTCAAGGTTTGCACAGGAGTGGGGGCTTCTATTGGTTAATTCAGGTTGTGATTTAGCCCG
          X
TTGGGTTTCATTAAGGTGTGTAAACAGGTGGGTGGTACCTGGAGGTTCATTCATTGGGATAACGAGAGGAGGAGGGGCTAGAGGTCCGCGAGATTGGGG
TTGGGTTACATTAAGGTGTGTAAACAGGTGGGTGGTACCTGGAGGTTCATTCATTGGGATAACGAGAGGAGGAGGGGCTAGAGGTCCGCGAGATTGGGG

TAGGCGGAGCTCAGGAGGGTCCCTCCATAGGGTTGAACGAGGAGGGGAGGATGGGCTCCGCCCGATATACCTAGTGGGTGGAGCTAGAGGTAGG
TAGGCGGAGCTCAGGAGGGTCCCTCCATAGGGTTGAACGAGGAGGGGAGGAGTGGGCTCCGCCCGATATACCTAGTGGGTGGAGCTAGAGGTAGG

TATCCATAGGGTCCATTATCTGGAGGTATCTAAGCTCCGCCCTATATACAGGTGGGTGGAGCTAGGTTAGGTTCAAGCTAGGTTCTTACGGGGTA
TCTCCATAGGGTCCATTATCTGGAGGTATCTAAGCTCCGCCCTATATACAGGTGGGTGGAGCTAGGTTAGGTTCAAGCTAGGTTCTTACGGGGTA

CCCCCTACCTACCTAAGGTGCGCCACCCTTCCCTCCCTTCCGTTTTAATGGTAGAATAACCTATAGGTTATTAACTAGTGGTGGAAATAGGGTATTGCA
CCCCCTACCTACCTAAGGTGCGCCACCCTTCCCTCCCTTCCGTTTTAATGGTAGAATAACCTATAGGTTATTAACTAGTGGTGGAAATAGGGTATTGCA

GCTGGGTATATACCTATAGGTATATAGAACCTAGAGGAAGGGAACCTATAGTGTATCCCTCCCCCTACCCCCCTCCCTACGGTTGCTGAGC
GCTGGGTATATACCTATAGGTATATAGAACCTAGAGGAAGGGAACCTATAGTGTATCCCTCCCCCTACCCCCCTCCCTACGGTTGCTGAGC

CCATCCCCACCCAGCACCCGGGGTGCCTGGCCCGCCCGCTTACTGACTTGTACCTTTGACATTTGGTCAGCTGACCGATGCTCGCCACTT
CCATCCCCACCCAGCACCCGGGGTGCCTGGCCCGCCCGCTTACTGACTTGTACCTTTGACATTTGGTCAGCTGACCGATGCTCGCCACTT

CCTGGGTATGACCTGGCTGTGCTTGTCCGTTGGCAATGTACCTCCAGCGTGGTGGCTTCTTTGGGATGCATCACITTAGGCCACTAAGCCCCGGT
CCTGGGTATGACCTGGCTGTGCTTGTCCGTTGGCAATGTCCCTCCAGCGTGGTGGCTTCTTTGGGATGCATCACITTAGGCCACTAAGCCCCGGT

TGCTCGCTTGGCTGCTCACCATGACACACTAAGCCCCTGTAAATCCATGAGCCCCGCTTTAGGAAGCACACGTCCTCCGGGACGGAAAGGGGACTTGG
TGCTCGCTTGGCTGCTCACCATGACACACTAAGCCCCTGTAAATCCATGAGCCCCGCTTTAGGAAGCACACGTCCTCCGGGACGGAAAG
    
```

B

```

CTGCGCCGCCGCAAGTCTTGGGGCAGCCGGGGTTCCTGGCGCTCCGGGGGACGCGGGCCGGCC-GCCGGTGGGTCCGTGGG-CCGCTGCCCGCTCCGGGTGGGGGGTGGCCCGCTGGGCA-CCG
-----GGTCC-GGGGT-TCCGG---CCCTGGAGCTCGGGGGGCGGCCGGGTGGCCACC---GGTCCGCTGGGTCCGTGCCCGCTCC-GGGGGGGTGGCCCGCTGGGCA-CCG
    
```

FIG. 4. (A) Alignment of the homologous regions D_L (upper line) and D_R (lower line). Identity of homologous bases are indicated by an asterisk. The initiation sites of the RNAs are marked by triangles, with small arrows showing the direction of transcription. The TATA-like sequences at positions -29 and -33 are boxed. Note that the noncoding strands of D_L and D_R are shown. X designates the position of an additional G in the D_R sequence. Different bases in the D_L sequence of B95-8 are shown above the M-ABA D_L sequence. The long arrows and the horizontal bars illustrate two inverted repeats and a palindrome of 10 bp, respectively. (B) Alignment of one *NotI* (upper line) and one *PstI* repeat unit (lower line). Symbols are the same as described above for those in panel A.

digestion of the DNA with *SacI*, creating two *SacI-KpnI* probes which differed in size by 173 bp (Fig. 5).

By hybridization with the D_R probe fragments of 222 and 223 bases were protected from S1 nuclease digestion (Fig. 5, lane 5). Two faint bands of 243 and 325 bases were also seen. The lower band represents a second RNA species with the initiation site 21 bases apart from the major transcript. The upper band could theoretically represent a third RNA species but can be more readily explained by an incomplete *PstI* digestion which increased the size of the probe by one repeat (102 bp).

One RNA from each region (D_L and D_R) is thus initiated at the identical position in the duplicated sequences. At 29 bp upstream of these cap sites a modified Goldberg-Hogness box (GATAAAA) could be identified. The second RNA started 21 bp further upstream in D_L near a second TATA

box-like element (TATTACA) 33 bp upstream of its cap site. In D_L the second promoter was used about two to three times more efficiently than the first. In D_R the second promoter was used almost exclusively. The different promoter usage in D_R probably reflects a single base pair change in the corresponding Goldberg-Hogness box.

A second region involved in transcription by RNA polymerase II is the canonical sequence CAAT located 70 to 90 bp upstream from the cap site of many eucaryotic RNAs (10). The motif of CAAT is found in the homologous regions D_L and D_R, 69 and 90 bp upstream of the cap sites, respectively.

Mapping of the 3' ends. The transcripts of D_L (2.5 kb) and D_R (2.8 kb) terminate in the left flanking sequences of the repeat clusters. The 3' ends were mapped by S1 nuclease analysis. DNA of the plasmid pM800 was digested with *BglI*,

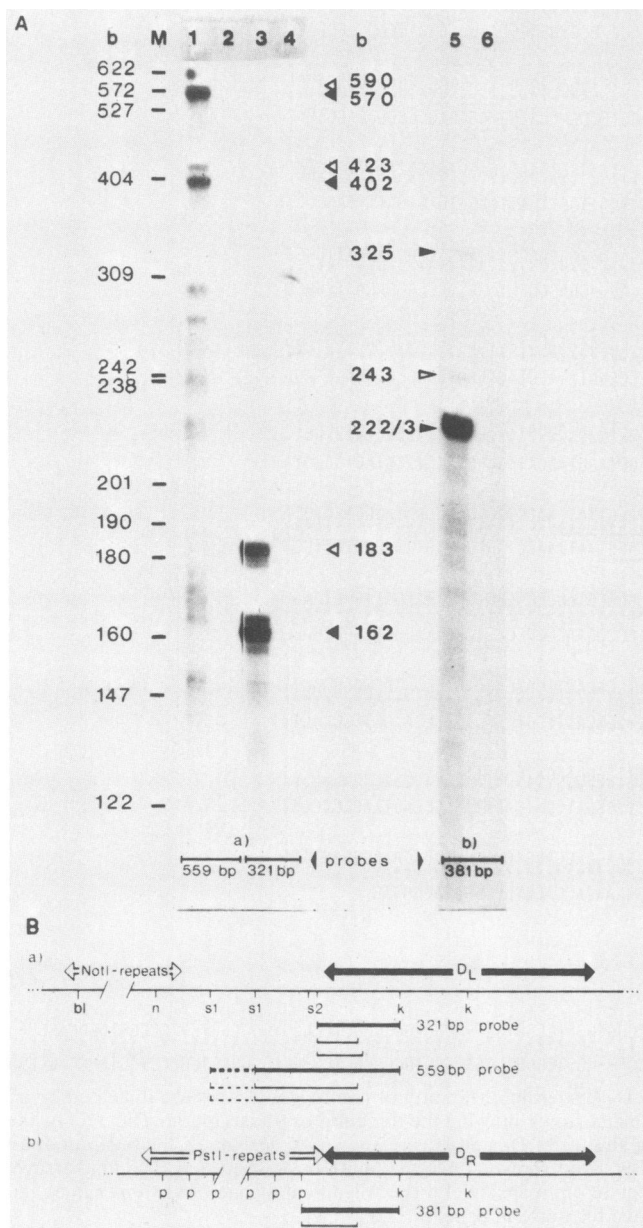


FIG. 5. (A) Autoradiography of the S1 nuclease protection experiment mapping the 5' ends. Cytoplasmic RNA of TPA-induced M-ABA cells (lanes 1, 3, and 5) and yeast tRNA (lanes 2, 4 and 6) were hybridized to the following probes: a *SacI-KpnI* fragment (lanes 1 and 2), a *SacII-KpnI* fragment (lanes 3 and 4), and a *PstI-KpnI* fragment (lanes 5 and 6). Samples were run on a denaturing 6% polyacrylamide gel after S1 nuclease digestion. The protected fragments are marked in bases. The RNAs start at two promoters (open and closed triangles) 21 bp apart from each other. The first promoter is less efficiently used in D_R (lane 5). A marker lane (M) is designated on the left side. (B) The probes are schematically shown by horizontal bars below the restriction enzyme maps (bl, *BglII*; k, *KpnI*; n, *NotI*; p, *PstI*; s1, *SacI*; s2, *SacII*) of the relevant regions. The parts protected from S1 nuclease digestion are shown as thin horizontal lines below the probes. Dashes indicate heterogeneity in the size of a probe due to partial digestion with *SacI*.

and DNA of pMB2-K was digested with either *PstI* or *SstI*. The antisense strands were labeled from their 3' ends with [α - 32 P]dCTP using T4 DNA polymerase. The labeled DNAs were cut with *BamHI* and *BglII*, respectively. The 699-bp *BamHI-BglII* (D_L), the 704-bp *BglII-PstI* (D_R), and the 632-bp *BglII-SstI* fragments (D_R) were isolated by preparative gel electrophoresis. These fragments spanned the boundaries between the *NotI* (D_L) and *PstI* repeats (D_R) and their left-flanking sequences. Fragments of 340 and 325 bases were protected from S1 nuclease digestion by hybridization of cytoplasmic RNA with the *BamHI-BglII* probe (Fig. 6). This indicates that the RNA is terminated at two distinct sites after the polyadenylation signals in unique sequences left of the *NotI* repeats 15 bases apart from each other.

By hybridization with the *BglII-PstI* and *BglII-SstI* probes, fragments of 155 and 79 bases, respectively, were protected (Fig. 6). An additional faint band was seen after hybridization with the *BglII-PstI* probe but not with the *BglII-SstI* probe, suggesting that it is generated by unspecific hybridization due to the high G+C content of the *PstI* repeats.

Results of the S1 nuclease protection experiments demonstrated that the RNAs from D_L and D_R were not spliced within the regions covered by the probes. We did not attempt to map the entire transcripts by S1 nuclease analysis because the clusters of tandem repeats will never form perfect duplexes and tend to anneal out of frame (1, 18).

Open reading frames. The fact that both transcripts are poly(A)⁺ and are localized at polyribosomes suggests that they are translated into polypeptides which might play a role in the lytic cycle of the virus.

The transcript starting in D_L has a long open reading frame beginning with an AUG in the spacer segment between D_L and the *NotI* repeats. The reading frame extends through the whole cluster of repeats. A large part of this hypothetical protein would be coded for by the *NotI* repeats. Since the number of base pairs per *NotI* repeat (125 bp) is not a multiple of 3, the reading frame would shift from one repeat unit to the other. Assuming that there are 12.3 *NotI* repeats would result in a protein with a molecular weight of about 70,000 with a repetitive structure of four times 125 amino acids within the molecule. Since the number of *NotI* repeats is variable in different EBV strains, different reading frames might be used for the C termini in different strains. It is remarkable that all three frames are terminated in front of the poly(A) addition site. The AUG in front of the long open reading frame is the second AUG with respect to the cap site. The reading frame starting at the first AUG codon would allow for a peptide of 35 amino acids, with the frame being stopped by a terminator codon (TGA). A second AUG codon can be used as the initiation site for protein synthesis if the first AUG codon is followed by a stop codon (23, 27). Only this second AUG codon fits the rule by Kozak that a purine residue should be at position -3 (22).

The transcript starting in D_R has a long open reading frame starting with an AUG codon at position 3302, extending through the whole *PstI* repeat cluster and terminating at the stop codon at position 614. This open reading frame could eventually code for a polypeptide with 896 amino acids and a molecular weight of 96,000. The protein would consist of a repeated structure of about 25 times 34 amino acids made up by the *PstI* repeats. The last *PstI* repeat at the 3' end of the repeat cluster is one nucleotide shorter than the average repeat units. Remarkably, because of this, the long open reading frame is terminated by a stop codon in front of the polyadenylation signal. The amino acid composition of both

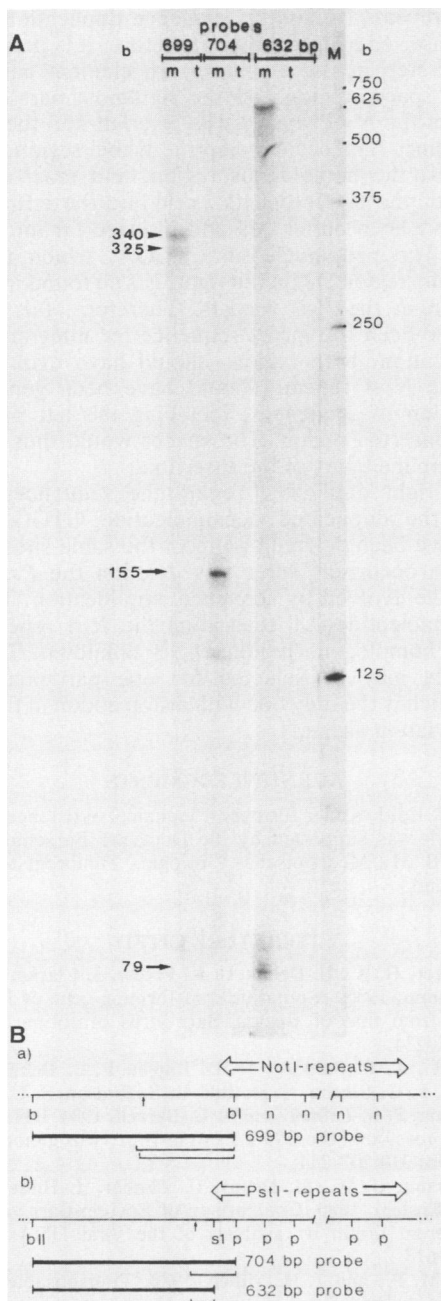


FIG. 6. (A) Autoradiography of the S1 nuclease protection experiment mapping the 3' ends. Cytoplasmic RNA of TPA-induced M-ABA cells (lanes m) and yeast tRNA (lanes t) were hybridized to the following probes: a *Bam*HI-*Bgl*II fragment (699 bp), a *Bgl*III-*Pst*I fragment (704 bp), and a *Bgl*III-*Sst*I fragment (632 bp). Samples were analyzed as for the mapping of the 5' ends. The protected fragments (in bases) are marked by arrows on the left side. A marker lane (M) is designated on the right side. (B) The probes are schematically shown by horizontal bars below the restriction enzyme maps (b, *Bam*HI; bl, *Bgl*II; bll, *Bgl*III; n, *Not*I; p, *Pst*I; and s1, *Sst*I) of the relevant regions. The parts protected from S1 nuclease digestion are shown as thin horizontal lines below the probes.

hypothetical proteins is shown in Table 1, which shows that there was a high proline and arginine content.

A number of proteins are known which have a repeated amino acid structure. Most of these, like collagen (24, 42),

TABLE 1. Amino acid composition of the hypothetical proteins coded by the regions carrying the *Not*I and *Pst*I repeats

Amino acid	Composition (%) of the region carrying the following repeats:	
	<i>Not</i> I	<i>Pst</i> I
Alanine	15.7	9.3
Arginine	14.1	17.5
Asparagine	1.0	2.8
Aspartic acid	2.2	0
Cysteine	2.6	2.9
Glutamine	5.5	8.6
Glutamic acid	2.3	2.8
Glycine	15.1	14.7
Histidine	2.3	0.2
Isoleucine	0	0
Leucine	3.8	3.0
Lysine	0	0.1
Methionine	0.3	0.1
Phenylalanine	0.1	0
Proline	21.3	23.1
Serine	5.7	3.2
Threonine	5.2	8.6
Tryptophan	1.0	2.9
Tyrosine	0.4	0
Valine	1.2	0.1

silk fibroin (28, 37), melting-point-lowering serum protein in arctic fish (43), and zein, the storage protein of maize (16), are composed of short repeat units of only a few amino acids. Two examples of repetitive proteins with somewhat larger repeat units have been described recently. The gene product of Balbiani ring 2 of *Chironomus tentans* is composed of a structure of about 70 amino acids repeated at least 25 times (39) and the circumsporozoite protein of *Plasmodium knowlesi* of a structure of 12 amino acids repeated 12 times (33). Compared with these proteins, the structure of the putative EBV polypeptides would still be remarkable.

The strange properties of the hypothetical gene products thus raises the question of whether the RNAs are indeed translated into protein. So far we have been unable to detect proteins of the respective size in various producer and nonproducer cells after TPA induction. In addition, RNA from TPA-induced Raji cells selected by hybridization to fragments with the repeat clusters did not reveal any in vitro translation products (19, 36a). Therefore, the possibility remains that these inducible transcripts are not translated and act as regulatory RNAs.

Polyadenylation signals. Termination of the D_L (2.5-kb) and D_R (2.8-kb) transcripts occur in the left-flanking sequences of the repeat clusters. The polyadenylation signal AATAAA (35) is present for both mRNAs (D_L and D_R, position 568, Fig. 3) at a distance of 260 and 53 nucleotides from the repeat clusters, respectively. The consensus sequence YGTGTTY has been found 15 to 35 bp downstream of herpes simplex virus (HSV) and other eucaryotic mRNAs (32). The sequence TGTGGTTT is found 24 and 30 bp downstream of the polyadenylation site in D_R and D_L, respectively. A similar sequence TGTGTTGT is additionally found 21 bp downstream of the poly(A) addition signal in D_L, possibly reflecting the second 3' end of the RNA shown in Fig. 6. The positions of these signals correspond to the overall lengths of both RNAs.

Sequence elements in the duplicated region compared with those in other eucaryotic genes. The region upstream from the cap sites of the RNAs in D_L and D_R was analyzed for

sequence elements also observed in other eucaryotic genes. Since EBV is a herpesvirus it is most obvious to make a comparison with the known genes of HSV. With regard to their regulation the HSV genes fall into three classes: immediate early (α), early (β), and late (γ) genes. For their own transcription β genes require the presence of a functional immediate early (α) gene product. The most prominent feature of β genes is an AC string 100 to 120 bp upstream from the cap site, which is not found in α or γ genes (6). The role of these elements in defining a gene as a β gene is not yet clear, however. Similar AC strings were also found 155 to 100 bp upstream from the cap site of the transcripts starting in D_L and D_R (Fig. 3).

An element with the sequence TAATGARAT is found upstream from HSV immediately early genes and is present in one to three copies at positions between -115 and -485 (29, 41). This element is distinguished from an enhancer element by its orientation dependence. Very similar motifs are found at positions -123 to -131, -329 to -337, and -102 to -110, -308 to -316 in the duplicated regions (Fig. 3).

In the rabbit β -globin gene another element has been identified at position -83 to -111 by site-directed mutagenesis and generation of deletion mutants, which is required for maximal transcription. This element consists of an imperfect tandem repeat of 14 to 15 bp. Comparative analysis of globin genes from different species has allowed the establishment of the consensus sequence CCNCACCCTG (9). Similar elements are also found in the HSV thymidine kinase gene and in the simian virus 40 early region.

In D_L and D_R an imperfect repeated element of 9 bp was found in four copies (position -152, -139, -106, and -77), with 1 bp exchange each, which were separated from each other by 4, 24, and 21 nucleotides, respectively. All four repeats are very close to the consensus sequence proposed by Dierks et al. (9). The last one was the same distance from the CAAT box as that observed in the human β -globin gene (6 bp).

Comparison of the upstream regions in D_L and D_R with those of the inducible genes of strain B95-8 (12, 13) did not reveal obvious similarities.

To analyze further the function of these regulatory elements it will be necessary to study their role in transcription after *in vitro* mutagenesis. Since both transcription units are silent in cells carrying EBV latently and are only transcribed on induction of a lytic or abortive cycle of the virus, these genes behave like β rather than α genes. This suggests that the sequence elements shared between HSV α genes and these transcription units do not represent enhancer elements and are involved in regulation of the activity of these genes in general. Using constructs of D_R and the chloramphenicol acetyl transferase gene, we are now attempting to identify the regions required for inducibility and optimal activity.

An evolutionary model for the generation of D_L and D_R . The finding that the repeat clusters are closely related and juxtaposed to almost completely conserved regions raises the questions of how they may have evolved. An evolutionary model also must account for the fact that the *NotI* repeats are separated from the homologous region D_L by 538 bp, whereas the *PstI* repeats from D_R are separated only by the dinucleotide AT. The starting point for an evolutionary model was the observation that a sequence located at the border from the spacer region to the homologous region D_L is present in inverted orientation within the *NotI* repeats (two mismatches within 14 nucleotides) bracketed by the decanucleotide GTGGGGGGTG in the same orientation.

Inverted repeats flanked by sequence duplications are features of inserted elements (4). Therefore, it is proposed that as a first step in the evolution, an element carrying the complete spacer region and the rightmost part of the sequences of the *NotI* repeats was inserted into the ancestral viral genome. As a second step the whole region, including the spacer, the homologous region, and most of the sequences of the *NotI* repeats excluding the leftmost part, would have been duplicated and inserted far into the viral genome. The pentanucleotide CTGGA, which flanks the homologous region D_R on the right, is also found in the same orientation in the *PstI* repeats. Therefore, this sequence could have been the target sequence for integration. After the duplication, both regions should have evolved differently. The *NotI* repeats should have been generated by amplification of sequences, including the left part of the formerly inserted region. The spacer would thus represent the nonamplified part of the insertion.

In the right duplicated region the sequences inserted between the duplicated decanucleotide GTGGGGGGTG should have been deleted at almost the same sites at which integration occurred. After this deletion the *PstI* repeats should have evolved by sequence amplification. The origin of the dinucleotide AT separating the *PstI* repeat cluster from the homologous region D_R is unknown. These two nucleotides may be selected for one particular reading frame, which is the only one with a stop codon in front of the polyadenylation signal.

ACKNOWLEDGMENTS

We thank Edith Kofler for expert technical assistance.

This work was supported by die Deutsche Forschungsgemeinschaft (SFB 31, Medizinische Virologie, Tumorentstehung und -Entwicklung).

LITERATURE CITED

1. Adldinger, H. K., H. Delius, U. K. Freese, J. Clarke, and G. W. Bornkamm. 1985. A putative transforming gene of Jijoye virus differs from that of Epstein-Barr virus prototypes. *Virology* 141:221-234.
2. Baer, R., A. T. Bankier, M. D. Biggin, P. L. Deininger, P. J. Farrell, T. G. Gibson, G. Hatfull, G. S. Hudson, S. C. Satchwell, C. Séguin, P. S. Tufnell, and B. G. Barrell. 1984. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. *Nature (London)* 310:207-211.
3. Bornkamm, G. W., H. Delius, U. Zimmer, J. Hudewentz, and M. A. Epstein. 1980. Comparison of Epstein-Barr virus strains of different origin by analysis of the viral DNAs. *J. Virol.* 35:603-618.
4. Calos, M. P., and J. H. Miller. 1980. Transposable elements. *Cell* 20:579-595.
5. Clewell, D. W., and D. R. Helinski. 1969. Supercoiled circular DNA-protein complex in *Escherichia coli*; purification and induced conversion to an open circular form. *Proc. Natl. Acad. Sci. USA* 62:1159-1166.
6. Costa, R. H., K. G. Draper, L. Banks, K. L. Powell, G. Cohen, R. Eisenberg, and E. K. Wagner. 1983. High-resolution characterization of herpes simplex virus type 1 transcripts encoding alkaline exonuclease and a 50,000-dalton protein tentatively identified as a capsid protein. *J. Virol.* 48:591-603.
7. Crawford, D. H., M. A. Epstein, G. W. Bornkamm, B. G. Achong, S. Finerty, and J. L. Thompson. 1979. Biological and biochemical observations on isolates of EB virus from the malignant epithelial cells of two nasopharyngeal carcinomas. *Int. J. Cancer* 24:294-302.
8. Dambaugh, T. R., and E. Kieff. 1982. Identification and nucleotide sequence of two similar tandem direct repeats in Epstein-Barr virus DNA. *J. Virol.* 44:823-833.
9. Dierks, P., A. van Royen, M. D. Cochran, C. Dobkin, J. Reiser,

- and C. Weissman. 1983. Three regions upstream from the cap site are required for efficient and accurate transcription of the rabbit β -globin gene in mouse 3T6 cells. *Cell* 32:695-706.
10. Efstratiadis, A., J. Posakony, T. Maniatis, R. Lawn, C. O'Connell, R. Spiritz, J. DeRiel, B. Forget, S. Weissman, J. Slightom, A. Blechl, O. Smithies, F. Baralle, C. Shoulders, and N. Proudfoot. 1980. The structure and evolution of the human β -globin gene family. *Cell* 21:653-668.
 11. Epstein, M. A., and B. G. Achong (ed.). 1979. The Epstein-Barr virus. Springer Verlag, Berlin.
 12. Farrell, P. J., A. Bankier, C. Séguin, P. Deininger, and B. Barrell. 1983. Latent and lytic cycle promoters of Epstein-Barr virus. *EMBO J.* 2:1331-1338.
 13. Farrell, P. J., P. L. Deininger, A. Bankier, and B. Barrell. 1983. Homologous upstream sequences near Epstein-Barr virus promoters. *Proc. Natl. Acad. Sci. USA* 80:1565-1569.
 14. Favalaro, J., R. Treisman, and R. Kamen. 1980. Transcription maps of polyoma virus-specific RNA: Analysis by two-dimensional S1 gel mapping. *Methods Enzymol.* 65:718-749.
 15. Freese, U. K., G. Laux, J. Hudewentz, E. Schwarz, and G. W. Bornkamm. 1983. Two distant clusters of partially homologous small repeats of Epstein-Barr virus are transcribed upon induction of an abortive or lytic cycle of the virus. *J. Virol.* 48:731-743.
 16. Geraghty, D., M. A. Peifer, I. Rubenstein, and J. Messing. 1981. The primary structure of a plant storage protein: zein. *Nucleic Acids Res.* 9:5163-5174.
 17. Hayward, S. D., S. G. Lazarowitz, and G. S. Hayward. 1982. Organization of the Epstein-Barr virus DNA molecule. II. Fine mapping of the boundaries of the internal repeat cluster of B95-8 and identification of additional small tandem repeats adjacent to the HR-1 deletion. *J. Virol.* 43:201-212.
 18. Hudewentz, J., H. Delius, U. K. Freese, U. Zimmer, and G. W. Bornkamm. 1982. Two distant regions of the Epstein-Barr virus genome with sequence homologies have the same orientation and involve small tandem repeats. *EMBO J.* 1:21-26.
 19. Hummel, M., and E. Kieff. 1982. Mapping of polypeptides encoded by the Epstein-Barr virus genome in productive infection. *Proc. Natl. Acad. Sci. USA* 79:5698-5702.
 20. Jeang, K. T., and S. D. Hayward. 1983. Organization of the Epstein-Barr virus DNA molecule. III. Location of the P3HR-1 deletion junction and characterization of the *NotI* repeat units that form part of the template for an abundant 12-*O*-tetradecanoyl-phorbol-13-acetate-induced mRNA transcript. *J. Virol.* 48:135-148.
 21. Jones, M. D., and B. E. Griffin. 1983. Clustered repeat sequences in the genome of Epstein-Barr virus. *Nucleic Acids Res.* 11:3919-3937.
 22. Kozak, M. 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.* 12:857-872.
 23. Kozak, M. 1984. Selection of initiation sites by eukaryotic ribosomes: effect of inserting AUG triplets upstream from the coding sequence for preproinsulin. *Nucleic Acids Res.* 12:3873-3893.
 24. Kramer, J. M., G. N. Cox, and D. Hirsh. 1982. Comparison of the complete sequences of two collagen genes from *Caenorhabditis elegans*. *Cell* 30:599-606.
 25. Kröger, M., and A. Kröger-Block. 1982. A flexible new computer program for handling DNA sequence data. *Nucleic Acids Res.* 10:229-236.
 26. Kröger, M., and A. Kröger-Block. 1984. Simplified computer programs for search of homology within nucleotide sequences. *Nucleic Acids Res.* 12:193-201.
 27. Liu, C.-C., C. C. Simonsen, and A. D. Levinson. 1984. Initiation of translation at internal AUG codons in mammalian cells. *Nature (London)* 309:82-85.
 28. Lizardi, P. M. 1979. Genetic polymorphism of silk fibroin studied by two-dimensional translation pause fingerprints. *Cell* 18:581-589.
 29. Mackem, S., and B. Roizman. 1982. Structural features of the herpes simplex virus α gene 4, 0, and 27 promoter-regulatory sequences which confer α regulation on chimeric thymidine kinase genes. *J. Virol.* 44:939-949.
 30. Maniatis, T., E. F. Fritsch, and J. Sambrook. 1982. Molecular Cloning. A laboratory manual, p. 117-121. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
 31. Maxam, A. M., and W. Gilbert. 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* 65:499-560.
 32. McLauchlan, J., and J. B. Clements. 1983. DNA sequence homology between two co-linear loci on the HSV genome which have different transforming abilities. *EMBO J.* 2:1953-1961.
 33. Ozaki, L. S., P. Svec, R. S. Nussenzweig, V. Nussenzweig, and G. N. Godson. 1983. Structure of the Plasmodium knowlesi gene coding for the circumsporozoite protein. *Cell* 34:815-822.
 34. Polack, A., G. Hartl, U. Zimmer, U. K. Freese, G. Laux, K. Takaki, B. Hohn, L. Gissmann, and G. W. Bornkamm. 1984. A complete set of overlapping cosmid clones and subclones from M-ABA virus, a transforming non-defective Epstein-Barr virus strain derived from a nasopharyngeal carcinoma: detailed restriction enzyme mapping reveals the close similarity to other virus isolates. *Gene* 27:279-288.
 35. Proudfoot, N. 1984. The end of the message and beyond. *Nature (London)* 307:412-413.
 36. Raab-Traub, N., T. Dambaugh, and E. Kieff. 1980. DNA of Epstein-Barr virus. VIII. B95-8, the previous prototype, is an unusual deletion derivative. *Cell* 22:257-267.
 - 36a. Seibl, R., and H. Wolf. 1985. Mapping of Epstein-Barr virus proteins on the genome by translation of hybrid-selected RNA from induced P3HR-1 cells and induced Raji cells. *Virology* 141:1-13.
 37. Sprague, K. V., M. B. Roth, R. F. Manning, and P. L. Gage. 1979. Alleles of the fibroin gene coding for proteins of different length. *Cell* 17:407-413.
 38. Staden, R. 1977. Sequence data handling by computer. *Nucleic Acids Res.* 4:4037-4051.
 39. Sümegi, J., L. Wieslander, and B. Daneholt. 1982. A hierarchic arrangement of the repetitive sequences in the Balbiani Ring 2 gene of *Chironomus tentans*. *Cell* 30:579-587.
 40. Vieira, J., and J. Messing. 1982. The pUC plasmids, an M12mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* 19:259-268.
 41. Whitton, J. L., and J. B. Clements. 1984. Replication origins and a sequence involved in coordinate inductions of the immediate-early gene family are conserved in an intergenic region of Herpes Simplex virus. *Nucleic Acids Res.* 12:2061-2079.
 42. Yamada, Y., V. E. Avvedimento, M. Mudryj, H. Ohkubo, G. Vogeli, M. Irani, I. Pastan, and B. de Crombrughe. 1980. The collagen genes: evidence for its evolutionary assembly by amplification of a DNA segment containing exon of 54 bp. *Cell* 22:887-892.
 43. Yčas, M. 1976. Origin of periodic proteins. *Fed. Proc.* 35: 2139-2140.