Journal of Cellular Biochemistry 19:267–274 (1982)
Tumor Viruses and Differentiation 49–56

Sequence Analysis and In Vitro Transcription of Portions of the Epstein-Barr Virus Genome

P. L. Deininger, A. Bankier, P. Farrell, R. Baer, and B. Barrell

MRC Laboratory of Molecular Biology, Cambridge CB22QH, England

The 17,180 base-pair Eco-RI-C fragment of Epstein-Barr virus has been sequenced in its entirety. This same fragment has also been analyzed for RNA polymerase II promoters, which are active in a soluble in vitro assay. These data are compared to the availability of predicted open reading frames and other potential nucleotide signals associated with transcription. In addition, the DNA sequence of a number of previously undetected repeated DNA sequences from this and several nearby regions of the viral genome are reported.

Key words: DNA sequence, repeated sequences, in vitro transcription

The Epstein-Barr virus (EBV) genome is made up of approximately 175,000 base pairs divided into two major unique sequence regions by a 3-kb internal, tandem repeat (see Fig. 1) [1]. The viral DNA also contains an approximately 500 base pair, tandem repeat at both ends, through which it is thought to circularize to an episomal form in the cell [2].

Epstein-Barr virus readily converts B lymphocytes to a growth-transformed state and exists as a latent infection. It is difficult to induce the virus to a high level of productive infection. There are three major poly-A-RNA species in transformed cells with latent virus [3,4] and approximately 50 species in productive infection [5]. All of these RNAs have been roughly mapped on the EBV genome, but their precise location has yet to be determined. Also, because of the low levels of expression, it is conceivable that minor RNA species or transiently expressed RNA species have been overlooked.

Epstein-Barr virus exists as a persistent infection in most adults and plays a causative role in infectious mononucleosis as well as a potential role in Burkitt's lymphoma and nasopharyngeal carcinoma [6]. To understand the mechanisms of EBV-induced diseases it is necessary to obtain precise knowledge of the structure and function of the EBV genome. Our initial approach in studying this virus is to

P.L. Deininger is now at the Department of Biochemistry, LSU Medical Center, 1901 Perdido St., New Orleans, LA 70112.

Received June 7, 1982; accepted July 1, 1982.

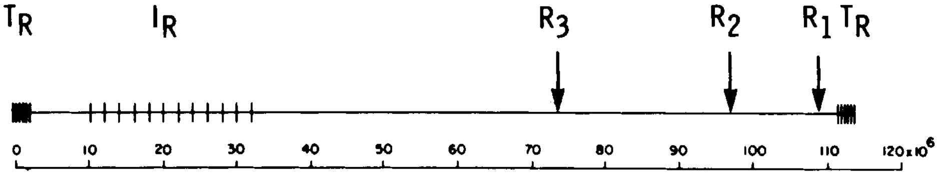

Fig. 1. The map of the Epstein-Barr virus genome. The genomic map represents approximately 175 kilobases of DNA organized as two major "unique" DNA regions separated by multiple copies of a 3-kb internal repeat (I_R). There are also multiple copies of an approximately 500-base repeat at both ends of the genome (T_R). We have found other repeated DNA regions, by sequence analysis, within the long unique regions (R_1, R_2, and R_3). There are almost certainly other repeated DNA regions not yet detected.

determine the nucleotide sequence of the entire virus. This approach has been made possible by recent advances in the DNA sequencing technology [7–9], as well as the availability of recombinant DNA libraries containing all of the viral DNA fragments [2]. The DNA sequence analysis is a preliminary characterization, which provides information on the genome structure and organization as well as providing a detailed map and subclone bank useful for detailed characterizations of genome expression.

We present here some of our findings from sequence analysis of a 17-kb restriction fragment, Eco-RI-C, from the B95-8 strain of EBV and preliminary studies to characterize its expression using in vitro RNA polymerase II transcription [10]. We also report repeated DNA sequence structures which had not been previously detected in this and adjoining regions of the EBV genome.

MATERIALS AND METHODS

The EBV clone library was supplied by B. Griffin and J. Arrand [2], and plasmid DNAs were prepared as described [2]. Recombinant DNA inserts were cleaved from the vector and isolated from low gel temperature agarose [11]. Sequencing was carried out using an M13 subcloning "shotgun" strategy (For 8) with subfragments created from sonicated DNA as has been described [9]. Data were handled by computer [12].

RNA polymerase II studies were carried out in vitro as previously described [10]. The restriction fragments for the transcription studies were predicted from the DNA sequence and isolated by electrophoresis in low gel temperature agarose [11].

RESULTS

The Eco-RI-C DNA Sequence

Sequence analysis was carried out predominately by "shotgun" sequencing of random fragments produced by sonication of the cloned Eco-RI-C fragment. The approach allowed the determination of the sequence of all but 70 nucleotides located around position 2500. The final 70 nucleotides were then determined by predicting the restriction map from the sequence and isolating a predicted Pst-1 to Mbo-1 restriction fragment spanning that region. At this point more than 90% of the sequence has been determined on both strands, and almost all the rest has been determined multiple times on independent subclones. This was felt to be necessary because the high G + C content (59%) leads to a relatively high level of artifacts in the sequencing gels and because there were no other data available to help corroborate the length and number of open reading frames. Even now it is difficult to consider the reading frames as unambiguous. The Eco-RI-C fragment contains 17,180 base pairs. For lack of space, we will present only selected portions of sequences and general conclusions.

The larger open reading frames predicted from the sequence are presented in Figure 2. This does not rule out the possibility that some smaller reading frames may actually be used by the virus, nor can we say that the virus definitely uses all of the reading frames presented. In conjunction with the reading frame map we also indicate all of the $^{5'}$AATAAA$^{3'}$ sequences in the fragment (Fig. 2). These sequences are found near the 3'ends of most mRNAs, and may indicate the approximate position of the 3' termini of some of the EBV RNA species.

We also find a previously undetected repeated sequence in the EBV genome, between nucleotides 8568 and 8893 on the Eco-RI-C fragment maps (Fig. 2; see also Fig. 1). The sequence of this repeated region and the relation of the homologous portions is indicated in Figure 4A.

In Vitro Transcription of the Eco-RI-C Fragment

In order to map potential RNA polymerase II promoters we used a soluble in vitro transcription assay system [10]. We find that very large fragments, such as the entire Eco-RI-C fragment, give weak signals in proportion to the background. We have therefore used a series of overlapping subfragments to cover the entire 17-kb region. The fragments used are shown in Figure 3 along with the map showing the location of promoters. We detected only three major promoters in this study. The direction and location of transcription was determined by comparing resulting transcription bands from overlapping fragments and were confirmed by a final transcription using replicative form DNA from M13 subclones carrying the predicted promoter. From these studies we cannot rule out the possibility that there were other promoters that were either too weak or totally inactive in the in vitro assay.

Sequence Analysis of Other Portions of the EBV Genome

Several other fragments of the EBV genome have been sequenced or are in the process of being sequenced in our laboratories. From these sequence studies several other previously undetected repeated DNA sequences can be detected in the EBV genome. The approximate locations of these repeated regions are marked in Figure 1. They are found in the EBV Eco-RI-D fragment and the Bam-HI-K fragment (Fig. 4A,B), which also contains the viral terminal repeat (10). The repeat in the Eco-RI-D region and its homology are presented in Figure 4C. We do not know its exact loca tion in the fragment because the DNA sequence analysis is not yet complete. The repeat in the Bam K fragment (R1, Fig. 1) is the most impressive in terms of its size and internal homology. A region of 685 bases contains only perfect copies of these closely related subunits (Fig. 4C).

DISCUSSION

Analysis of Open Reading Frames

The sequence of the EBV Eco-RI-C fragment allows us to predict the locations and size of the major open reading frames in this region of the genome. Figure 3 shows the 14 largest open reading frames in this 17,180-nucleotide fragment of DNA.

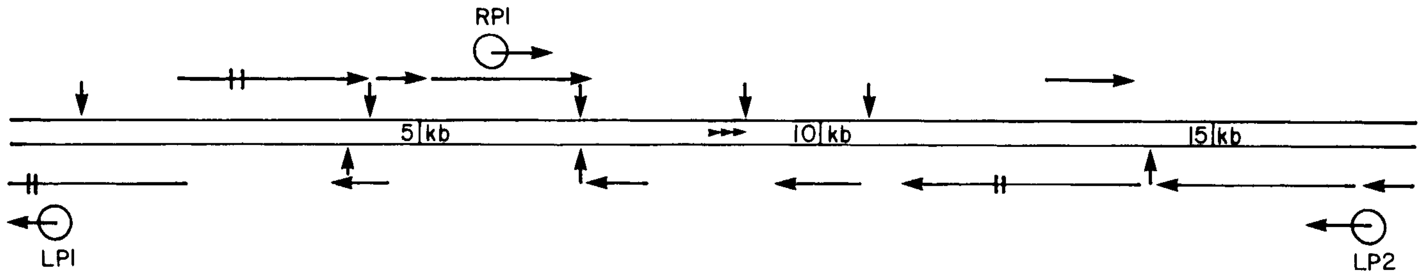

Fig. 2. Reading frame map of the EBV Eco-RI-C fragment. The 17,180 base pair DNA sequence was translated into amino acid sequence in all three reading frames in both directions. Any reading frame greater than 500 nucleotides in length is indicated by an arrow in the appropriate direction of translation. Reading frames interrupted by a pair of vertical lines actually represent two reading frames that overlap in the region between the lines. Vertical arrows represent the sequence $^{5'}$AATAAA$^{3'}$. If the arrow is above the map, the AATAAA is read from left to right. If the arrow is below the map, the sequence is present on the other strand. The approximate map position of sequences that act as RNA polymerase II promoters in vitro is represented by a circle with an arrow to represent the direction of transcription. These promoters are designated LP1, LP2, and RP1. The three horizontal arrowheads placed together at about position 8500 mark the location of a repetitive sequence described in Figure 4A.

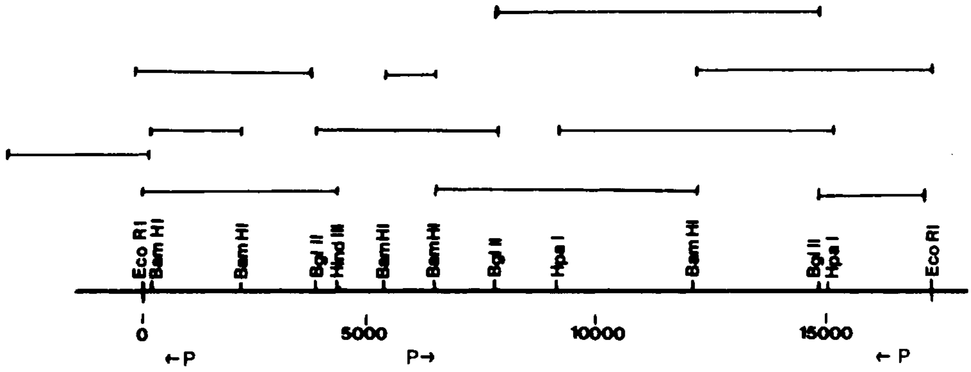

Fig. 3. Map of fragments used to map RNA polymerase II promoters. A series of subfragments were isolated from low melting point agarose after digestion of the Eco-RI-C fragment with the appropriate enzymes. Overlapping fragments were prepared to make sure no active promoters were overlooked, and the result of the in vitro polymerase II assay is shown in Figure 2.

Two of these open reading frames extend into the adjoining EBV DNA fragments, and the others range in size from approximately 500 to 2,500 nucleotides. In a lytic EBV infection there are approximately ten transcripts detected from this region of the genome ranging from 1.1 kb to 6.4 kb [5]. This suggests the possibility of RNA splicing events to create many of the RNA species. Another possibility is that errors in our DNA sequence may lead to frame shifts and that some of the open reading frames should actually be fused. In a project of this size and technical difficulty, it is difficult to rule out this possibility completely.

The presence of the sequence $^{5'}$AATAAA$^{3'}$at the 3′termini of most messenger RNAs is useful to help define the ends of RNA species. Several copies of this sequence present in this fragment are located at or near the end of open reading frames (Fig. 3). It is clear that all reading frames do not terminate with this sequence nearby. Thus, in order to contain this sequence, messenger RNAs encoding some of the reading frames would have to splice to other reading frames with an AATAAA. This sequence is also found several times far from the ends of open reading frames (Fig. 2). We do not know whether the unusual location of these sequences is significant. In a genome of high G + C content there should be even more selection against the random presence of such an A+T-rich sequence than there is in a typical genome, which should help select against the random occurrence of AATAAA.

One very unusual feature of these reading frames is that around positions 13,000 to 14,000 and 4,700 to 4,000 there are sizable open reading frames on both DNA strands. The use of two such reading frames is unprecedented, and they may not both be used by the virus. Unused reading frames of this size are also very rare, as noncod ing DNA usually contains frequent termination codons. Since termination codons are A+T-rich, it is possible that a G+C-rich genome, like EBV, will contain fewer terminators, making larger unused reading frames more likely. Except for these two examples, however, the strand opposite a large open reading frame in the sequence is terminated frequently.

```
A    TTTGTTAATCTTTAGTGGGAACTAGTGGGAGTGCTGTGCCTC
      ||||||| |||| |||| | ||||||  |||| |||||
     ATTGTTAACCTTTGGTGGAACCTAGTGTTAGTGTTGTGCTGTAAATAAGTGTCCAGCGCACCACT
      |   |||| |||||||| ||||  || |  || ||||||||||||||||| | ||| || |||
     CTCTGTAACATTTGGTGGGACCTGATGCTGCTGGTGTGCTGTAAATAAGTGCCTAGCACATCACG
```

```
B    TGACAATGGCCCACAGGACCCTGACAACACTGA

     TGACAATGGCCCACAGGACCCTGACAACACTGA

     TGACAATGGCCCACATGACCCGCT
                    -     ---

              GCCTCAGGACCCTGACAACACTGA
              -  -

     TGACAATGGCCCACAGGACCCTGACAACACTGA

     TGACAATGGCCCACATGACCCGCT
                    -     ---

              GCCTCATAGCCCT
              -  -  ---
```

```
C      -Y-X-X-Z-Z-Z-X-X-Z-Z-X-X-Z-X-Z-Z-X-X-Z-X-Z-Z-X-Y-Z-Z-

     Y-X-Z-Z-Y-Y-Y-Y-Z-Y-Z-Y-Z-X-Z-X-Z-Y-Z-Y-Z-Y-Z-X-Y-Y-
```

X = GGGGCAGGAGCAGGAGGA

Y = GCGGCAGGAGCAGGA

Z = GGGGCAGGA

Fig. 4. Minor repetitive sequences in the EBV genome. The sequences of three minor repetitive sequences found within the EBV genome are presented. These sequences are located as shown in Figure 1. A) This sequence (R_2 in Fig. 1) is located within the Eco-RI-C fragment (see Fig. 2). Homology between the different subunits is shown by a vertical line. B) A sequence located in the Eco-RI-D fragment, near the terminal repeat (R_1 in Fig. 1). Positions of nonhomology are underlined. C) A repeat consisting of only three triplets—GGA, GGG, and GCA—located in the Bam-HI-K fragment (R_3 in Fig. 1). These triplets may also be organized into the three larger units shown, X, Y, and Z, which are then arranged together apparently randomly.

Almost all of the Eco-RI-C fragment is covered by open reading frames on one strand or the other. There is a region of about 1.5 kb that has no apparent coding regions. This region contains a repeated DNA sequence which will be discussed below. Also, the B95-8 strain of the virus has a 15-kb deletion relative to most other EBV isolates [13] that maps within this apparently noncoding region (B. Barrell, unpublished observation). The initial deletion may have led to some alterations in the region that have destroyed reading frames, although restriction mapping has shown no major evidence of this [13].

In Vitro Transcription of the Eco-R1-C Fragment

We have used a soluble extract system to map the RNA polymerase II promoters active in vitro. We find only three major promoters active in this system (Fig. 3).One promoter initiates synthesis just before the open reading frame which runs out the left end of the Eco-RI-C fragment. Another promoter near the right end of the fragment initiates just before one of the largest open reading frames facing to the left. In close agreement with the present mapping data [5], this reading frame is also punctuated by an AATAAA sequence near its termination codon, suggesting that there may be a simple, unspliced message from this region of about 3 kb. The third promoter appears to initiate within an open reading frame. At the present time we have no explanation for this.

Although the in vitro RNA polymerase II system does not appear to detect false promoters in eucaryotic DNA, it also does not seem to recognize all eucaryotic promoters. It is possible that there are a number of other promoters in this region that are very weak or silent in the soluble extracts. It is thus difficult to assess the extent of splicing that may be needed to use the available open reading frames, although our data suggest that splicing is common.

Repeated DNA Sequences in the EBV Genome

The B95-8 strain of Epstein-Barr virus is known to contain two major repeated DNA regions: One is the internal repeated DNA region, which divides the 15-kb short unique region from the 120-kb long unique region [1–3]. This repeat unit con tains approximately ten copies of a 3-kb long sequence. At both ends of the viral genome are also approximately eight copies of a 0.5-kb sequence on each end [1,2,13]. Both the internal and the terminal repeats vary in copy number in different virus particles.

Our DNA sequence studies have detected a number of smaller repeated DNA regions within the long unique region. These smaller repeated DNA sequences are found within the Eco-RI-C fragment (R_2, Figs. 1 and 2), in the Eco-RI-D-Het fragment about 1,600 nucleotides from the terminal repeat and in the Bam-K fragment, which is in the middle of the long unique region. Thus, these different repeated DNA sequences are found in widely differing locations in the EBV genome. These repeats essentially subdivide the long unique region of the EBV genome into a series of smaller unique fragments.

The repeated DNA sequence in the Eco-RI-C fragment is the most highly diverged of the repeated DNA sequences discovered (Fig. 4). There are essentially three copies of a sequence approximately 65 bases long, with a 23-base deletion in one copy. The average homology is only 75%. It is located in a region of the genome containing no open reading frames, and is near the location of the 15-kb deletion in this strain.

The repeat detected in the Eco-RI-D fragment consists of 184 bases arranged with a basic repeat unit of 33 bases (R_1, Figs. 1 and 4B). However, there are two sets of deletions that occur at three repeat unit intervals, which suggests this repeat originated in several independent duplication steps, much as has been described for many eucaryotic satellite DNA sequences [14].

The repeat found in the Bam-HI-K fragment l (R_3, Fig. 1) is perhaps the most striking of these sequences (Fig. 4C). That sequence of 705 bases is made of only three triplets, GGG, GCA, and GGA, organized into related 9 mers, 15 mers, and 18 mers, which are then placed together in a seemingly random pattern. This sequence contains no translation terminators in any of the six reading frames. A sequence of this size with an unused open reading frame is unusual, and with six open reading frames even more so. The Bam-K fragment is a region of major transcriptional activity in transformed cells [3,4], but if this repeat were part of the transcriptional unit it would code for a very unusual, repetitive protein.

ACKNOWLEDGMENTS

Prescott Deininger was supported by a fellowship from the North Atlantic Treaty Organization. P.D. and R.B. would like to thank Dr. Frederic Sanger for space in his laboratory and helpful discussions.

REFERENCES

1. Rymo N, Forsblum S: Nucleic Acids Res 5:1387, 1978.
2. Arrand J, Rymo L, Walsh J, Bjorck E, Lindahl T, Griffin B: Nucleic Acids Res 9:2999, 1980.
3. King W, Thomas-Powell A, Raab-Traub N, Hawke M, Kieff E: J Virol 36:506, 1981.
4. King W, Van Santen V, Kieff E: J Virol 38:649, 1981.
5. Hummel M, Kieff E: J Virol 43:262, 1982.
6. Epstein M, Achong B: "The Epstein-Barr Virus," Berlin: Springer Verlag, 1979, pp 1–38.
7. Sanger F, Coulson A, Barrell B, Smith A, Roe B: J Mol Biol 143:161, 1980.
8. Messing J, Crea LR, Seeburg P: Nuclieic Acids Res 9:309, 1981.
9. Fuhrman S, Deininger P, LaPorte P, Friedmann T, Geiduschek E: Nucleic Acids Res 9:6439, 1981.
10. Manley J, Fire A, Cano A, Sharp P, Gefter M: Proc Natl Acad Sci USA 77:3855, 1980.
11. Wieslander L: Anal Biochem 98:305, 1979.
12. Staden R: Nucleic Acids Res 8:3673, 1980.
13. Heller M, Dambaugh T, Kieff E: J Virol 38:632, 1981.
14. Deininger P, Jolly D, Rubin C, Friedmann T, Schmid C: J Mol Biol 151:17, 1981.