Clustered repeat sequences in the genome of Epstein Barr virus

M.D.Jones⁺ and Beverly E.Griffin

Imperial Cancer Research Fund, Lincoln's Inn Fields, London WC2A 3PX, UK

Received 8 April 1983; Accepted 10 May 1983

ABSTRACT

The genome of Epstein-Barr virus is composed of unique DNA interspersed with repetitive sequences. This organisation suggests that Epstein-Barr virus provides a useful model for studying the function(s) of repetitive sequences in eukaryotic chromosomes. The primary structure of two of the repeat sequences, the 3072 bp large internal repeat, or BamHI-W repeat, and a smaller 125 bp, G,C-rich NotI repeat, are presented here. Their structures and possible functions are discussed.

INTRODUCTION

One of the more interesting and unresolved problems arising from studies on mammalian cellular DNA concerns the nature and function of the many repeating sequences found within individual species. The presence and expression of repetitive sequences in eukaryotic DNA, broadly separated into three classes that have been designed unique (about one copy per haploid genome), moderately repetitive (10^3 - 10^5 copies), and highly repetitive (10^6 or greater), has been reviewed recently by Jelinek and Schmid (1). In the well-characterised small DNA tumour viruses, such as adeno or papova viruses, such repeats seem to be absent, whereas they occur in herpesviruses. In this regard, the genome of the human herpesvirus, Epstein-Barr virus (EBV), appears to resemble mammalian chromosomal DNA in that it also contains a considerable amount of repetitive DNA. For example, Rymo and Forsblom (2) showed that B95-8 virion EBV DNA contains about eleven copies of an internal BamHI restriction fragment, roughly 2.0 megadaltons in size. Subsequent to this, the same, or very similar, repeating sequence, now termed the large internal repeat sequence, has been found in every strain of EBV

examined, in episomal forms of the EBV genome, and in chromosomal DNA from EBV-transformed cells of Burkitt lymphoma, nasopharyngeal carcinoma, and infectious mononucleosis origins. Recent studies have shown the presence of many other, smaller, repeating sequences within EBV DNA (3), such as have been described previously for herpes simplex viruses (4). One of these, about 0.1 megadaltons in size, is present also in roughly eleven copies and has been isolated as a NotI restriction enzyme fragment (Hayward et al., 1982). Both this and the large internal repeat are present in tandem arrays of sequence. Thus, EBV would appear to provide an excellent model system for studying the function of clustered repeats present in cellular chromosomes. In this paper, we present the sequences of the two repeats isolated as BamHI and NotI restriction fragments and discuss their structures and potential for coding.

MATERIALS AND METHODS

Cloning into M13 Vectors. DNA from recombinant plasmids containing EBV restriction fragments BamHI-W, -C, -Y, -H and EcoRI-A were prepared as described previously (5). The M13 vectors mp7, mp8 and mp9, and the host strain JM101 (6,7) were generous gifts of A.T. Bankier and B.G. Barrell. Recombinant plasmid DNA was digested with the appropriate restriction enzymes and the resulting fragments isolated by electrophoresis and subsequent elution from low temperature gelling agarose, as described by Sanger et al., (8). Suitable fragments were cloned into an M13 vector.

Random clones from the above were obtained by the ultrasonic shearing method described by Fuhrman et al. (9). Briefly, an isolated EBV restriction fragment, for example BamHI-W, was ligated in the presence of T4 DNA ligase to form circular or multimeric species, and subsequently degraded by sonication. The termini of degraded DNA molecules were repaired by treatment with T4 DNA polymerase in the presence of all four deoxynucleoside triphosphates and the products were separated by agarose gel electrophoresis. DNA fragments 300-800 base pairs long, as assessed by comparison with DNA standards, were isolated and cloned into blunt-ended M13 vectors. The vectors

used were either M13mp7 or M13mp8, cleaved with HincII or SmaI, respectively. Subsequent transformation of E.coli JM101 and isolation of single-stranded DNA from recombinant clones was carried out as described by Sanger et al. (8).

In addition to the above, specific fragments were cloned and isolated for use in DNA sequence determinations: DNA of the recombinant clone containing the EBV restriction fragment EcoRI-A (5) was digested with BglII and the resulting fragments separated by agarose gel electrophoresis. The BglII repeat fragment was isolated, cloned into BamHI cleaved pAT153 and transfected into E.coli HB101. A recombinant clone containing the BglII repeat unit was isolated and DNA obtained according to the procedure of Birnboim and Doly (10). DNA from this clone was digested with HinfI, the ends were "repaired", and fragments 350-400 pairs in length were isolated and cloned into a M13 vector, as above.

DNA of the recombinant clone containing EBV BamHI-fragment H was digested with the enzyme NotI, treated with T4 DNA polymerase and ligated directly into the SmaI site of M13mp8. Using a similar procedure, the small SstII-PvuII sub-fragment of BamHI-W was cloned directly into M13mp8.

These experiments were all carried out under GMAG Category 1 and 2 conditions.

DNA Sequencing. The sequences of M13 recombinant clones were determined using the dideoxy chain termination method of Sanger et al., (8, 11), and a chemically-synthesised heptadecanucleotide primer (12). DNA sequence data were handled using the programmes BATIN, DBCOMP and DBUTIL (13). Analysis of the data was performed using SEQ (14). Nucleotide residues 1 to 84 and 2935 to 3072 from the large internal repeat were also determined by the chemical degradation method of Maxam and Gilbert (15).

RESULTS

Sequence of the BamHI-W, Large Internal Repeat, Fragment. The restriction fragment, BamHI-W (Fig. 1), has been shown to occur approximately eleven times in the B95-8 isolate of Epstein-Barr virion DNA (2). We have cloned this restriction fragment, as well as the rest of the EBV genome, as reported previously (5)

Figure 1. A schematic diagram that illustrates the strategy for sequencing the BamHI-W (large internal repeat) fragment of EBV DNA. A single cloned species (designated clone number B16-4H) was fragmented by sonication and recloned in M13 as described in Materials and Methods. Individual clones were analysed using the dideoxy method of Sanger et al., (8, 11). Each arrow represents a region of the DNA whose sequence was determined by this method, either in a left to right direction (+) or the reverse (-). Because the cloning procedure is non-specific, many duplicate sequences were obtained, and only the minimum number used to obtain the entire sequence are illustrated. Overall, about 80 sequencing gels, each providing data on 200-300 nucleotides were analysed. The symbols shown represent: ★, data obtained using the Maxam-Gilbert (15) sequencing method; ▲, data obtained from HinfI sub-fragments of BglII versions of the large internal repeat; ■, data obtained from HinfI fragments that allow sequences which transverse the BamHI junctions to be determined. The dashed arrow (----) shows the location of an SstII-PvuII sub-fragment which was cloned and its sequence determined in order to resolve a discrepancy between our data and those of Cheung and Kieff (18). Additional data are described in the text.

Some restriction enzyme sites derived from the sequence are illustrated.

and in order to examine the nature and structure of this highly conserved, repetitive, and thereby presumably important, region, we have determined its DNA sequence.

The sequencing strategy adopted utilised "shotgun" cloning into the phage M13, and the Sanger dideoxy sequencing methodology (6, 8, 11). An isolate of BamHI-W was ligated to produce multimeric forms, which were then sonicated. The resulting DNA fragments were treated with T4 DNA polymerase to repair the ends and then fractionated into different size classes by electrophoresis on an agarose gel. Fragments between 300 and 800 base pairs were isolated and ligated into blunt-ended M13

vectors (either HincII or SmaI cleaved M13 mp7 or mp8). Colourless plaques were picked, the recombinant phage grown, and DNA isolated and sequenced (see Fig. 1). From this random approach, almost 90% of the DNA sequence was obtained. The remaining gaps in the sequence data were obtained by 'directed' or 'forced' cloning of relevant restriction enzyme fragments (7). The data were ordered using the DBUTIL programme of Staden (13). The sequence of the 3072 base-pair long BamHI-W fragment is presented in Figs. 2 and 3. The strategy adopted for sequencing yielded about 14,000 base pairs of sequence information, which covered over 85% of the sequence on both strands. Only 2% (66bp) of the sequence has been determined using single clones, and these have been re-sequenced a number of times under varying conditions, and analysed on gels of different compositions. The overall data obtained agree well with the estimated molecular weight of about 2×10^6 daltons for this large internal repeat sequence (2, 16, 17). The sequence differs at two sites from that recently published by Cheung and Kieff (18) and contains one nucleotide more than these authors reported, that is, 3072 rather than 3071 base pairs.

In order to check that no small restriction fragment was located between BamHI repeat fragments, a BglII version of the large internal repeat was obtained from a cosmid clone that contained the entire B95-8 EBV EcoRI-A restriction fragment (5). It was cloned into the plasmid vector pAT153 (19) and a HinfI restriction fragment, spanning the BamHI junction between BamHI-W repeats (see Fig. 1), was isolated, cloned into M13mp8 and its sequence determined. The data (not given) showed that the large BamHI-W fragments are tandemly repeated, and no other small BamHI fragment exists between them.

Location of the Junctions Between Unique and Repeat Sequences.

The sequence origins of the large internal repeat region lie within BamHI restriction fragments C and Y (20, 21). The DNA sequences of these sites were also determined. Mapping data on the BamHI-C fragment (J.R. Arrand, unpublished) has shown that a TaqI-PvuII restriction fragment contains the junction between the large internal repeat sequence (LIRS) and a unique region of

18	20	34	40	50	60	70	80	90	100	110	120
GGATCCCGCC	ACCGGCCCTT	CTCTCTGTCC	CTCCACACCC	TAGACCCGAG	CTTCTGGCGT	CCCCGGGTCC	ACCAAGCCAG	CCGGAGGAGC	CGCGGAGGAC	CGCGCGAGCC	
GGATCCCGCC	ACCGGCCCTT	CTCTCTGTCC	CTCCACACCC	TAGACCCGAG	CTTCTGGCGT	CCCCGGGTCC	ACCAAGCCAG	CCGGAGGAGC	CGCGGAGGAC	CGCGCGAGCC	
130	140	150	160	170	180	190	200	210	220	230	
GGGCGGAGTC	TCCCTTGCCC	TCCTCTGCCC	CGAGCCCGCT	CAGTCTGCGT	CCCCACCGGC	TCCATCAGCC	GGGCGGAGG	GACCCGCGGC	GACCCGCGGC	GCGCGGTGTC	
GGGCGGAGTC	TCCCTTGCCC	TCCTCTGCCC	CGAGCCCGCT	CAGTCTGCGT	CCCCACCGGC	TCCATCAGCC	GGGCGGAGG	GACCCGCGGC	GACCCGCGGC	GCGCGGTGTC	
250	260	270	280	290	300	310	320	330	340	350	
AGTCTCTGCC	GCAGCGGCC	TCGACGACAG	GGGCGGACAT	TTTCTCCGCC	CAGCTGTAGC	CCGAGTCGAG	GGGTGAGAG	ACAGTGACA	GACATTTTAC	GTCGAGAGCT	
AGTCTCTGCC	GCAGCGGCC	TCGACGACAG	GGGCGGACAT	TTTCTCCGCC	CAGCTGTAGC	CCGAGTCGAG	GGGTGAGAG	ACAGTGACA	GACATTTTAC	GTCGAGAGCT	
370	380	390	400	410	420	430	440	450	460	470	
GGGCTCTCCGC	GGCCAGCCCT	CCCGACGGTG	GACAGGCCAT	GACAGGCCAT	GGGCGGAGG	CCGCGTGGG	CCGCGTGGG	CCGCGTGGG	CCGCGTGGG	CCGCGTGGG	
GGGCTCTCCGC	GGCCAGCCCT	CCCGACGGTG	GACAGGCCAT	GACAGGCCAT	GGGCGGAGG	CCGCGTGGG	CCGCGTGGG	CCGCGTGGG	CCGCGTGGG	CCGCGTGGG	
430	440	450	520	530	540	550	560	570	580	590	
GGCTCTCCGC	TAGCCTGCCC	TCGAGTAGGC	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
GGCTCTCCGC	TAGCCTGCCC	TCGAGTAGGC	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
610	620	630	640	650	660	670	680	690	700	710	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
730	740	750	760	770	780	790	800	810	820	830	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
850	860	870	880	890	900	910	920	930	940	950	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
970	980	990	1000	1010	1020	1030	1040	1050	1060	1070	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
1090	1100	1110	1120	1130	1140	1150	1160	1170	1180	1190	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
1210	1220	1230	1240	1250	1260	1270	1280	1290	1300	1310	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
1330	1340	1350	1360	1370	1380	1390	1400	1410	1420	1430	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
1450	1460	1470	1480	1490	1500	1510	1520	1530	1540	1550	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
1570	1580	1590	1600	1610	1620	1630	1640	1650	1660	1670	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
1690	1700	1710	1720	1730	1740	1750	1760	1770	1780	1790	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	
GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	GGGCGGAGG	

1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920
CTCCACAGA	CCCGGCTTC	AGGCTCTGC	TCCTCTGTA	CCCTTTTGA	ACCAAGATG	GGACACATG	GCCAGAGCC	TTGGCTCTA	AGGCTCTGC	GTCCCTCTGC	ACCCCGGCTT
GAGGTCTCT	GGGCGCGAG	TCCGGACGG	AGAGACAAT	GGGAAATCT	TGGTGTGGA	CTGTGTGTA	CGGTCTCGG	MACCGGAGT	TCCGGGAGC	CAGGGGAGC	TGGGGCGGA
1930	1940	1950	1960	1970	1980	1990	2000	2010	2020	2030	2040
CAGCAACCT	CTCTCTCCC	TCTCTCCAC	CCAGCTTCC	CCCTCTCCG	TCCCTCTCG	CTCTGTATC	TCCCTCCGT	CCCAAGTGG	CCGCTGCCC	CCCTGACCC	AGTACTGCG
GTCTGTGGA	CGACGAGGG	AGACGGTGG	GGTCGAGGG	GGGAGGGGG	AGGGGAGAG	GAGGACTAG	AGGGGGCCG	GGGTATCCC	GGGGACGGG	GGGACGTGG	TCATGGACG
2050	2060	2070	2080	2090	2100	2110	2120	2130	2140	2150	2160
CTCTTGGCC	ACGCACCCG	GGCGAGGCA	CGTTAGACC	GGCCAGCCC	CATCCCTGA	GACCCAGGG	CCATCTCTC	TGGTACGAG	CAGAGAAGA	GTAGAGGCC	CGCGCATTTG
GGAAACCGG	TGCTGGGGC	CCGGTCCGT	GGAATCTGG	CCGGTCTGG	GTAGGAGATT	CTGGGTGCC	GGTAGAGAG	ACCATGCTC	GTCTCTCTT	CATCTCGGG	GGCGGTATC
2170	2180	2190	2200	2210	2220	2230	2240	2250	2260	2270	2280
GGCCACAGT	GAGAGACCA	TCCAGGGCC	CCAGATTGA	CCAGAGGCC	ACCCAGAGT	CCAGACCCG	GTCTCTCCG	GGGGCAGAG	CAGGCAGGG	CCCCCGGAG	CTGGCCGCCA
CGGGCTTAA	CTCTCTGTC	AGTCTCCGG	GTCCACACT	CGTCCGCCG	TGGGTCTAG	GGTCTGGAG	CAGGAGGCC	CCCCCTCTT	GTCTCTCTG	GGGGCGGTC	GACCGGGCT
2290	2300	2310	2320	2330	2340	2350	2360	2370	2380	2390	2400
GGAGCGGCC	GAGTGGGCG	CGTGGGCTG	GGCTGGGCG	CGCCGGGCT	GGGAGTCTG	GGGTGGGCG	CTCTCTCTC	CAGAGGGCC	CTGGTCTCG	CGGTGGCCC	TGGGGTAAGT
CTCTCGGGG	CTCACCAGC	GGCAAGGAC	CGACCGGCT	CGGCGGCTC	CGGCCGAGA	CCACCCGCT	GGACGAGAG	GTCTCTCCC	GACCCAGGG	GCCACCGGG	ACCCCTATC
2410	2420	2430	2440	2450	2460	2470	2480	2490	2500	2510	2520
CTGGAGGCA	GAGGTGGCG	CTAGCCCGG	GGAGTGGCG	GGGATCGCC	CGGGTCTCT	TTGGCAGAG	CCGGGCGATC	CTCTGAGAC	CTCCGGGCC	GGACGGTCC	CTCAGCGCCC
GACCTCCGT	CTCCACAGC	GATCCGGCC	CTTCACCTC	CGGCTAGCG	GGCCAGAG	AGCCGTCTA	GGCCGCGCT	GAGACTCTG	GAGGCCCGG	CTGCGCAGG	GAGTCTGGG
2530	2540	2550	2560	2570	2580	2590	2600	2610	2620	2630	2640
CCAGACAGC	CCCAGGTCT	CCAGGAGGG	TCCGGCATC	TCAGGGGCG	CAGGCTCAC	ACCACAGCC	CCCCAGACC	GGGTCTCGC	CAGCCAGCC	GACCGGCCC	CGGCTTGGG
GGTCTGTCT	GGGTCTCAG	GGTCTCCAG	GGTCTCCAG	AGTCCCGCT	GTCCGAGTG	TGGTCTCGG	GGGTCTGGG	CCCAGAGCC	GTCCGCTCG	CTGCGCGGG	CGCGGACCG
2650	2660	2670	2680	2690	2700	2710	2720	2730	2740	2750	2760
CTCTCTGCG	CCGAGCGCC	GGGGTTGGT	CTGCCCCCT	CTCTGTCTT	CAGAGAAC	AGGGACCTC	GGCACCCAG	AGCCCTCTG	GGCCGCTCC	AGGGGCCCC	CTGGTCTCG
GGAGGAGCC	CGGTGCGCG	CCCCAACAA	GACGGGAGA	GAGACAGGA	GTCTCTTGG	TCCTCTGAG	CCGTGGGGT	TCGGGGAGC	CGGGCGGAG	TCCGCGGAG	GACCCAGGC
2770	2780	2790	2800	2810	2820	2830	2840	2850	2860	2870	2880
CTCCCTCTG	AGCCCGGTA	AACCCAAAG	ATGCTGAGG	GGAGCCACC	TCGGGGGCG	GGCCAGAG	TCCAGAGTC	AGGGACCTC	CAGGTGCTC	CCCGGGTCC	CAGCCAGCC
GAGGGAGAC	TCGGGCAAI	TTGGGTTCT	TACAGATCC	CTCTGGTGG	AGCCCGGCT	CGGGGCTCT	AGGTCTCCAG	TCCCTGTGA	GTCCACAGG	GGGGCCAGG	GTCCGCTCG
2890	2900	2910	2920	2930	2940	2950	2960	2970	2980	2990	3000
GAGAGAGAC	GGGACACCT	GGGCGCCCA	GAGGCGGCT	CTCGGCCCT	TCAGAGGCT	TCAGAGGCT	GGATGCTCC	CAGAGGGAG	CCTAGGCGT	CCCTCTCTC	CCCTCAGCC
CTCTCTCTG	GGGCTGCGG	CCGCGGGGT	CCTCCGCGA	AGGCGGGGA	AGGCGGGGA	AGTCTCGGT	CTACAGGGG	GTCTCTCTG	GGATCCGAG	GGGAGGAGG	GGGAGGTCC
3010	3020	3030	3040	3050	3060	3070	3080	3090	3100	3110	3120
CCGAGCTCT	CCCTCAGGA	GAGGCGGCT	TTTGGGCTC	CAATGCTAG	CCACAGGAG	CCGAGTGGC	CC				
GGCTCGGTA	GGGAGCGCT	CTCCCGGAG	AAACCCGGA	GTTCAGTCC	GGGTGCTCT	GGGCTCACG	GG				

Figure 2. The complete nucleotide sequence of a 3072 base-pair BamHI-W (large internal repeat) fragment of EBV DNA. The first six nucleotides shown make up the BamHI (GGATCC) restriction enzyme recognition site. Details of this sequence, which has an overall G + C content of 66.9%, as compared with 58-59% for the entire genome (43, 44) are discussed in the text.

the EBV genome (Figure 4). This fragment was isolated, cloned into M13mp8 and mp9 vectors and its DNA sequence determined. The junction between the LIRS and unique sequence was found to occur (leftward) between nucleotides 1858 and 1859 (Fig. 2). The sequences of the LIRS, starting from the unique end of the TaqI-PvuII sub-fragment of BamHI-C was identical to the corresponding part of BamHI-W (residues 1859-1896) except at residue 1882, where in BamHI-W the residue is T and in BamHI-C it is G (data not shown). Note that this replacement does not introduce a termination codon into the sequence. From the DNA sequence of the 1.9 kbp BamHI-Y fragment (our unpublished data), the junction between the LIRS and a unique (rightward) region of EBV in BamHI-Y has been determined. This occurs between residues 657 and 658 in the LIRS. Our finding agrees with that previously published (20). Thus, BamHI fragments C and Y contain 1213 and 657 bp of the LIRS, respectively, within them. Consequently, B95-8 EBV DNA contains an apparently non-integer number of large internal repeats (see Discussion). The amount of the tandemly repetitive (LIRS) BamHI-W sequence in the flanking restriction fragments C and Y, as determined by sequence analysis, corresponds to 0.61 of an intact LIRS, a

Figure 3. Polyacrylamide DNA sequencing gels which show data that differ from those reported elsewhere (18) for the large internal repeat, or correct our own reported sequence (35). **A.** A 6% polyacrylamide, 8M urea gel that shows the presence of four consecutive cytosine (C) residues between nucleotides 2627 and 2630 (Fig. 2). Cheung and Kieff (18) report three residues at position 2626 in their sequence. **B,C.** Gels (10% polyacrylamide, 8M urea) which illustrate how errors in DNA sequencing can be made, particularly in molecules with high G, C contents. A sequence that appears to be AGGA in four different sequencing reactions on one strand (two of which are shown, see gel B) of the DNA appears (see gel C) as TCCCT on the other strand. This discrepancy occurs at position 2233 on the sequence shown in Fig. 2, or at 2231 in the sequence reported previously (18).

result which is in excellent agreement with that obtained by Hayward et al. (21) using restriction enzyme analysis.

Homology of BamHI-W Repeat Sequences. In scanning for micro-heterogeneity among repeating sequences, several different BamHI-W clones were subjected to exhaustive analysis with a wide variety of restriction enzymes (M.D. Jones and A. Harris, unpublished). No sequence differences among the LIRS, using this approach, were observed. Moreover, Hayward et al. (17) have shown with intact EBV DNA that restriction patterns of the LIRS are identical. These data support the concept that the BamHI-W clone whose sequence has been determined (Fig. 2) is representative of other copies of the large repetitive sequence.

Coding Capacity of the Large Internal Repeat Sequence. We have analysed the coding capacity of the LIRS and the results are shown in Figs. 4 and 5. In brief, there is one particularly long open reading frame that proceeds in a left to right direction on the conventional BamHI restriction enzyme map of EBV DNA (22); this lies between nucleotides 2399 and 476 (Fig. 2). Elsewhere in the sequence, there are regulatory transcriptional signals, that is, a 'CAT' box (position 1100, Fig. 2) and a 'TATA' box (position 1138, Fig. 2, see Fig. 5) which are mutually located as described for other mammalian cells (23, 24). Noteworthy is the fact that the first ATG initiation codon subsequent to these signals, at position 1195 (Fig. 2), is followed immediately by a TAA termination codon. The DNA sequence therefore suggests that the BamHI-W repeating region is not within itself a complete coding entity. In support of this notion, there is no polyadenylation signal (AAT AAA) within the entire sequence. The presence of the long (1149 base pairs) open reading frame (top line, Fig. 5) and the overall coding potential (Fig. 5), would nonetheless appear to confirm the transcription data (25) which suggest that some part of this sequence must be transcribed, and, deducing from the sequence, could be translated.

DNA Sequence of a Small Internal (NotI) Repeat. Hayward et al. (21) and Hudewentz et al. (3) have shown that in addition to the BamHI-W sequence, a small cluster of about 11 copies of a tandem repeat sequence is present in the BamHI-H fragment, whose basic

Figure 5. A schematic diagram showing potential coding frames within the large internal repeat of EBV DNA, as deduced from the DNA sequence. A single repeating unit is shown here, with the nucleotide numbers taken from the sequence given in Fig. 2. Since there are multiple copies of the repeat sequence, for accuracy, this figure should be considered as if it were circular. Restriction enzyme sites, as given, allow correlation to be made between a single repeat unit, as shown here, and the Bam-W fragment, see Fig. 4.

At the top, the three reading frames are analysed from left to right, proceeding from BamHI-W toward BamHI-Y on a standard physical map of EBV DNA (see Fig. 4a), and at the bottom, the three frames analysed from the opposite direction, that is, proceeding from BamHI-W toward BamHI-C. Single vertical black lines show the location of termination codons. Potential extensive coding frames are shown as horizontal black bars with potential initiation codons (AUG) inserted as white bars. The topmost black bar, subsequent to a termination codon around position 2400, see below, corresponds to the open reading frame shown in Fig. 4b and 4c. (Initiation codons lying outside large potential coding frames are not indicated. Moreover, some areas designated as open reading frames would only specify a protein were the single termination codon, indicated, removed by appropriate splicing of the respective RNA).

As illustrated at the bottom of this figure, most of the potential coding sequences within a single BamHI-W fragment have counterparts in either BamHI-C or -Y, as shown (Fig. 4). If the concept that BamHI-W may be divided into two functionally different repeat sequences were correct, that portion which has been designated as potentially coding (C + Y) is very G, C-rich (>72%), whereas that designated as being derived from "W only" has the same overall composition (58%) as total EBV DNA. Moreover, the latter contains a high proportion of termination codons as shown.

```

      10      20      30      40      50      60      70
GGCGGCCGCC GGTGGGTCCG CTGGGCCGCT GCCCCGCTCC GGGTGGGGGG TGGCCCCGCT GGGCACCCT
CCGCCGCCG CCACCCAGGC GACCCGCGCA CGGGGCGAGG CCCACCCCCC ACCGGGGCGA CCCGTGGCGA

      80      90     100     110     120     125
GCGCCGCCGC CAGGTCTTGG GGCAGCCGGG GTTCCTGGCG CTCCGGGGGC AGCCG
CGCGGGCGGC GTCCAGGACC CCGTCGGCCC CAAGGACCGC GAGCCCCCG TCGGC

```

Figure 6. The complete nucleotide sequence of a 125 base-pair NotI repeat sequence of EBV DNA, with the first ten nucleotides containing the enzyme recognition site (GGCGGCCGCC). The NotI fragments are sub-fragments of BamHI-H (21). A single HaeII site (GGCGCT, underlined) lies between nucleotides 107-112.

repeating fragments into M13mp8 and determined their DNA sequence. Five recombinant clones have been sequenced in both orientations. Sequence data from all five are identical. Their sequence, which is 125 bp long and has a G+C content of 84.8%, is presented in Fig. 6. Our sequence data agree with revised data of Jeang et al. (27). One of the sequence origins of the NotI repeat has been determined during the course of analysis of the BamHI-H fragment (our unpublished data). This occurs at residue 115 (Fig. 6). Although there are no termination codons in these repeats (and no transcriptional signals), the fact that the sequence is so C,G-rich, coupled with the fact that it is not a multiple of three, makes it unlikely that it codes for a protein.

DISCUSSION

The discovery of high copy numbers of repetitive sequences within chromosomes of eukaryotic cells was made fifteen years ago (28). Documentation in support of this finding comes from numerous sources, and many putative functions have been put forward. (For reviews, see refs. 1, 29-31.) The obvious major hindrance of functional definition and understanding of these sequences, which account for a large proportion of the DNA within a cell, is the large size of the eukaryotic chromosome. For this reason, we would argue that the EBV genome presents an ideal model system for studies in this field. Perhaps more than any other virus described to date, it appears to "mimic" its host in being composed of a complex mixture of unique and repetitive sequences. In this paper, we present the

DNA structure of two quite different repeat sequences, one, the large (3072 bp) BamHI-W repeat, and the other the small (125 bp) NotI repeat*.

EB virions contain linear, double-stranded DNA molecules, about 170 Kb in size. A number of repeat sequences have been identified within the EBV genome, including terminal (32, 33) and several internal repeats (2, 3, 16, 21; B.G. Barrell and R. Baer, personal communication). One of the interesting observations on the EBV repeat sequences studied in detail to date is that they appear to contain a high percentage of guanine and cytosine. If this proves generally true, it seems likely that the denaturation maps, described by Bornkamm et al. (34) for DNAs from various strains of EBV essentially represent maps of the repetitive sequences within these species.

One of the interesting properties of tandem repeat sequences is that, in general, they appear to be structurally precise (2). For reasons discussed below, we feel this may also be the case in the two sequences under consideration here (Figs 2 and 6). In the light of three differences that were found between our initial BamHI-W sequence (35) and the 3071 bp sequence reported by Cheung and Kieff (18) which would contradict this concept, and the very important fact that these differences lie within a potential coding region (see Figs 4 and 5), we were obliged to re-investigate our data. Two of the differences are best considered by examination of the sequencing gels shown in Fig.3. In Fig. 3A can be seen the four consecutive cytidines which we report at nucleotide position 2627 (top strand, Fig. 2).

Sequencing the other strand of the DNA (data not shown) gave severe nucleotide compressions in this region (see ref. 36). This might explain why Cheung and Kieff found only three cytidine residues at this position. Further, in our original sequence, we defined position 2233-2234 as consisting of GG (bottom strand, Fig. 2), whereas Cheung and Kieff report three consecutive guanines at this site. Figs. 3B and 3C show how our error was initially made. The third difference involves a substitution by Cheung and Kieff to T for G at position 2505, relative to our sequence. This may represent a genuine polymorphism, or be due, more trivially, to the misreading of a DNA

sequencing gel. The overall net sequence of the BamHI-W fragment appears to consist of 3072 base pairs. Since the inconsistencies between our data (Fig. 2, and see above) and those of Cheung and Kieff (18) could be accounted for by sequencing errors, all BamHI-W repeats may be identical. It is unlikely that both of the determined sequences are derived from the same cloned species. Similarly, since we have analysed the primary structure of numerous clones of the NotI repeat and the data show them to be identical, and to agree with those of Jeang et al. (27), we think this repeat consists of 125 base pairs (see Fig. 6) and is also precise, despite data to the contrary; that is, Dambaugh and Kieff (26) report a 124 base pair repeat, the G-C pair at position 53 being absent in their sequence.

The BamH-W sequence (see Fig. 5) can be seen to consist of two quite distinct regions. In the portion that is derived in part from the BamHI-C and Y fragments (illustrated in Fig. 4) relatively few termination codons exist. That portion unique to BamHI-W, on the other hand, contains many termination codons in all six potential reading frames. In addition, it contains the transcriptional signals, that is, 'TATA and CAT' boxes, normally associated with eukaryotic chromosomes (23, 24). These facts suggest an alternative way of considering this repeat sequence. That is, instead of viewing in toto as a non-integer number of repeats, tandemly arrayed, it may functionally represent two repeating interdispersed elements, one (that comprised of sequences present in BamHI-C and Y) occurring in twelve copies in B95-8 DNA and the other (from BamHI-W only), in eleven copies. If the former is part of a coding sequence for a protein, the latter may control the number of copies that are actually translated. Thus, BamHI-W repeat could be representative of those families of sequences which consist of "genes" and "pseudogenes". In this connection, Lemischka and Sharp (37) have recently analysed a functional and pseudo- α -tubulin gene and found the latter to have a dispersed repetitive element within it.

Cells transformed by EBV have been shown to contain a mRNA, about 3 kilobases in size, which has sequence homology with the large internal repeat (25). Any theoretical consideration of

the BamHI-W as a coding sequence is complicated, however, by the paucity of potential initiation codons and the frequent presence of termination codons adjacent to or near putative initiation sequences (see text). This favours the notion that the transcriptional signals found within this fragment may not be associated in coding terms with the region that immediately follows them, but rather with some other part of the genome. The intriguing possibility exists that a promoter within any one of the repeat sequences can be used to generate mRNAs with different sizes of "leader" sequences, such as are found among the transcripts of the late genes of the small DNA tumour viruses, polyoma and SV40 (for discussion, see ref. 38). It is possible that the potential coding sequence present in BamHI-W may have its control signals elsewhere in the genome, for example in BamHI-C. These hypotheses can be examined once the sequences surrounding BamHI-W become available.

The overall structure of the large internal repeat contains a number of smaller repetitive sequences, as well as dyad symmetries which could potentially give rise to interesting secondary structures. Many of these have been discussed by Cheung and Kieff (18) and are unaffected by changes we propose in the sequence. However, these authors draw an analogy between a sixteen base pair sequence within the repeat and regions designated as "origins of replication" in the small DNA tumour viruses SV40 and BKV. This analogy is probably not valid since the observed homology covers only a part of the region that gives rise to the putative hairpin loop secondary structure which apparently is important in the replication of these viruses. Moreover, it does not include the ubiquitous A, T-rich sequence present in these viruses, as well as in polyoma virus, also probably essential for replication (see ref. 39). In addition, recent studies on autonomously replicating sequences (ars) in EBV do not implicate this structure (40). Thus, we feel that the nucleotide sequence per se provides little evidence that BamHI-W is encoding an origin of replication. Nor, in spite of potential secondary structure (not shown) within the NotI repeat sequence, is there any evidence, even circumstantial, that it could specify an origin of replication. It is also

unlikely that the latter contains coding information, not simply because it is so G, C-rich that it would encode a very unusual protein, but because coding would require proceeding from one frame to another along the repeats, yielding ultimately an unusual "tripartite" protein.

Whereas the primary sequence of the BamHI-W fragment suggests that it has coding potential, the most likely function (if any) for the portion of the genome covered by the NotI repeats would be to create some overall irregular or unusual structure within the DNA; that is, for example, to convert it to one or other of the alternative conformations of double-stranded DNA, such as Z-form DNA or "bent" forms of B-DNA (see refs. 41, 42). If so, this area of the genome may be of considerable biological significance. In this context, it is noteworthy that this region shows homology to another part of the EBV genome (3).

ACKNOWLEDGEMENTS

We are extremely grateful to A.T. Bankier, B.G. Barrell and P.L. Deininger for strains and advice on M13 cloning and sequencing. We thank J.R. Arrand for mapping data and advice on cloning the BamHI-C junction fragment, T.M. Sheedy and H.E. Ruley for help with the chemical sequencing, and J.R. Arrand and T. Lindahl for many useful suggestions and Ms S. Somani for help in preparation of manuscript.

⁺Current address: Department of Chemistry, Imperial College of Science and Technology, London SW7, UK

* Since DNA sequence studies are providing that the genome of this virus is scattered with repeat sequences (B.G. Barrell and R. Baer, personal communication), we prefer not to use the nomenclature (IR1, IR2, etc.) adopted by others (see ref. 26), since we believe it will soon prove awkward. Instead, we here define repetitive sequences in terms of restriction enzymes which allow them to be cleaved intact from the genome. Thus, the large internal repeat (LIRS) is referred to as a BamHI-W repeat, and the smaller repeat, within BamHI-H (see Fig. 1), as the NotI repeat.

REFERENCES

1. Jelinek, W.R. and Schmid, C.W. (1982) *Ann. Rev. Biochem.* 51, 813-844.
2. Rymo, L. and Forsblom, S. (1978) *Nucleic Acids Res.* 5, 1387-1402.
3. Hudewentz, J., Delius, H., Freese, U.K., Zimmer, U. and Bornkamm, G.W. (1982) *The EMBO Journal* 1, 21-26.
4. Davidson, A.J. and Wilkie, N.M. (1981) *J. Gen. Virol.* 55, 315-331.
5. Arrand, J.R., Rymo, L., Walsh, J.E., Bjorck, E., Lindahl, T. and Griffin, B.E. (1981) *Nucleic Acids Res.* 9, 2999-3014.
6. Messing, J., Crea, R. and Seeburg, P.H. (1981) *Nucleic Acids Res.* 9, 309-321.
7. Messing, J. and Vieira, J. (1982) *Gene* 19, 269-276.
8. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A. (1980) *J. Mol. Biol.* 143, 161-178.
9. Fuhrman, S.A., Deininger, P.L., LaPorte, P., Friedmann, T. and Geiduschek, E.P. (1981) *Nucleic Acids Res.* 9, 6439-6456.
10. Birnboim, H.C., and Doly, J. (1979) *Nucleic Acids Res.* 7, 1513-1523.
11. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
12. Duckworth, M.L., Gait, M.J., Goelet, P., Hong, G.F., Singh, M. and Titmas, R.C. (1981) *Nucleic Acids Res.* 6, 1691-1706.
13. Staden, R., (1980) *Nucleic Acids Res.* 8, 817-825.
14. Brutlag, D.L., Clayton, J., Friedland, P. and Kedes, L.H. (1982) *Nucleic Acids Res.* 10, 279-294.
15. Maxam, A.M. and Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* 74, 560-564.
16. Given, D. and Kieff, E. (1979) *J. Virol.* 31, 315-324.
17. Hayward, S.D., Noguee, L. and Hayward, G.S. (1980). *J. Virol.* 33, 507-521.
18. Cheung, A. and Kieff, E. (1982) *J. Virol.* 44, 286-294.
19. Twigg, A.J. and Sherratt, D. (1980) *Nature* 283, 216-218.
20. Cheung, A. and Kieff, E. (1981) *J. Virol.* 40, 501-507.
21. Hayward, S.D., Lazarowitz, S.G. and Hayward, G.S. (1982) *J. Virol.* 43, 201-212.
22. Skare, J. and Strominger, J.L. (1980) *Proc. Natl. Acad. Sci. USA* 77, 3860-3864.
23. Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) *Nucleic Acids Res.* 8, 127-142.
24. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeReil, J.K., Forget, B.E., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) *Cell* 21, 653-668.
25. van Santen, V., Cheung, A. and Kieff, E. (1981) *Proc. Natl. Acad. Sci. USA* 78, 1930-1934.
26. Dambaugh, T.R. and Kieff, E. (1982) *J. Virol.* 44, 823-833.
27. Jeang, K.-T., Hayward, G.S. and Hayward, S.D., Personal Communication.
28. Britten and Kohne (1968) *Science* 161, 529-540.
29. Brutlag, D.L. (1980) *Ann. Rev. Genet.* 14, 121-144.
30. Lewin, R. (1982) *Science* 217, 621-623.
31. Ulla, E. (1982). *TIBS*, June, 216-219.

32. Given, D., Yee, D., Griem, K. and Kieff, E. (1979) J. Virol. 30, 852-862.
33. Kintner, C.R. and Sugden, B. (1979) Cell 17, 661-671.
34. Bornkamm, G.W., Delius, H., Zimmer, U., Hudewentz, J. and Epstein, M.A. (1980) J. Virol. 35, 603-618.
35. Jones, M.D., Arrand, J.R., Sheedy, T., Karran, L., Harris, A. and Griffin, B.E. (1982) Abstract, Sixth Cold Spring Harbor Herpes Meeting, p.43.
36. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Petersen, G.B. (1982). J. Mol. Biol. 162, 729-773.
37. Lemischka, I. and Sharp, P.A. (1982) Nature 300, 330-335.
38. Acheson, N. (1980) In, "DNA Tumor Viruses" (ed. J. Tooze). Cold Spring Harbor Lab., pp. 163-185.
39. Soeda, E., Arrand, J.R., Smolar, N. and Griffin, B.E. (1979) Cell 17, 357-370.
40. Henry, B.E., Raab-Traub, N.J. and Pagano, J.S. (1983) Proc. Natl. Acad. Sci. USA 80, 1096-1100.
41. Zimmerman, S. (1982) Ann. Rev. Biochem. 51, 395-427.
42. Marini, J.C., Levene, S.D., Crothers, D.M. and England, P.T. (1982) Proc. Natl. Acad. Sci. USA 79, 7664-7668.
43. Schulte-Holthausen, H. and zur Hausen, H. (1970) Virology 170, 776-779.
44. Jehn, U., Lindahl, T. and Klein, G. (1972) J. Gen. Virol. 16, 409-411.