JOURNAL OF VIROLOGY, Oct. 1982, p. 286–294
0022-538X/82/100286-09$02.00/0

Vol. 44, No. 1

Long Internal Direct Repeat in Epstein-Barr Virus DNA

ANDREW CHEUNG AND ELLIOTT KIEFF*

Departments of Medicine and Microbiology, The University of Chicago, Chicago, Illinois 60637

Received 7 April 1982/Accepted 10 June 1982

The nucleotide sequence of the long internal reiteration, IR1, of Epstein-Barr Virus DNA has been determined. The repeat unit is 3,071 base pairs which are 66.8% guanine plus cytosine. There is a CCAAT sequence 39 nucleotides 5′ to a TATAA, which could indicate a promotor for transcription. The longest open reading frame is 1,124 base pairs. Also within IR1 is a sequence homologous to the papovavirus origin of DNA replication. The *ori*-like sequence is within a long palindromic region which is 500 base pairs overall. The palidromic region shares common features with the *Alu* family members and with eucaryotic transposable elements. The juncture between the short unique region (U1) and IR1 is also sequenced. The transition occurs in *Bam*HI-C at 1,214 base pairs before the *Bam*HI site in the first repeat of IR1. The transition from IR1 to the rightward unique region (U2) has been reported to be at 636 base pairs after the *Bam*HI site in the last repeat of IR1. Thus, relative to the start of IR1 at the juncture with U1, the last copy of IR1 is a partial repeat which contains only the beginning 1,850 base pairs.

Five classes of direct repeats are now recognized in Epstein-Barr virus (EBV) DNA (Fig. 1A). A variable number of direct repeats of a 500-base pair (bp) sequence, TR, are directly repeated at both ends of the genome (6–8, 10, 12). Four classes of internal direct repeats divide the genome into five unique sequence domains. The overall structure of the genome can be presented as TR U1 IR1 U2 IR2 U3 IR3 U4 IR4 U5 TR. IR1, IR2, and IR4 are tandem repeats of 3,000 (3, 10, 11, 16, 34), 123 and 103 (T. Dambaugh and E. Kieff, J. Virol., in press) bp units, respectively. IR3 is approximately 700 bp overall and differs from IR1, IR2, and IR4 in that it is composed of only three different nucleotide triplets (20a).

Each copy of the IR1 repeat has a single *Bam*HI site (5, 11, 16, 34). The repeat unit, *Bam*HI-V, the juncture fragment *Bam*HI-C, bounded by the last *Bam*HI site in U1 and the first *Bam*HI site in IR1, and the juncture fragment *Bam*HI-X, bounded by the last *Bam*HI site in IR1 and the first *Bam*HI site in U2 (Fig. 1A), have been cloned into pBR322 and are part of a library of recombinant EBV-pBR322 DNAs (5). The terminal *Hin*fI fragments of *Bam*HI-V have been sequenced and contain 35-bp direct repeats (3). The juncture between IR1 and U2 has also been sequenced and is 636 nucleotides after the last *Bam*HI site in IR1 (3). The extent of IR1 in the *Bam*HI-C juncture fragment is less than the length of the large *Sst*I fragment of *Bam*HI-V (3). The juncture between U1 and IR1 is therefore less than 2,200 nucleotides before the *Bam*HI site in IR1. Since IR1 begins less than 2,200 nucleotides before the first *Bam*HI site and ends 636 nucleotides after the last *Bam*HI site, relative to the start of IR1 at the juncture with U1, IR1 joins U2 in a partial copy which is at least 150 nucleotides short of being complete (3).

IR1 is highly conserved as a repeat element and as a homologous nucleotide sequence among the genomes of EBV and the primate viruses, herpesvirus papio and herpesvirus pan, which are genetically related to EBV (6, 18, 19, 20, 25). The conservation of IR1 suggests that it is an important DNA sequence. One function of IR1 is to encode part of a 3-kilobase (kb) cytoplasmic polyadenylated RNA in latently infected, growth transformed cells (22, 40). IR1 also encodes cytoplasmic polyadenylated and polyribosomal RNA in latently infected Burkitt tumor cells (7, 23, 30).

IR1 encodes very few cytoplasmic 3-kb RNA molecules per latently infected, growth-transformed cell (40). Direct analysis of the RNA is therefore difficult. Sequencing of the genomic DNA was undertaken in an effort to develop additional information about the possible functional characteristics of this region. In this manuscript, we report the nucleotide sequence of IR1 and of the U1-IR1 juncture.

MATERIALS AND METHODS

Viral DNAs. Plasmids pDK10 (*Bam*HI-C), pDK51 (*Bam*HI-V), and pDK14 (*Bam*HI-V), which contain

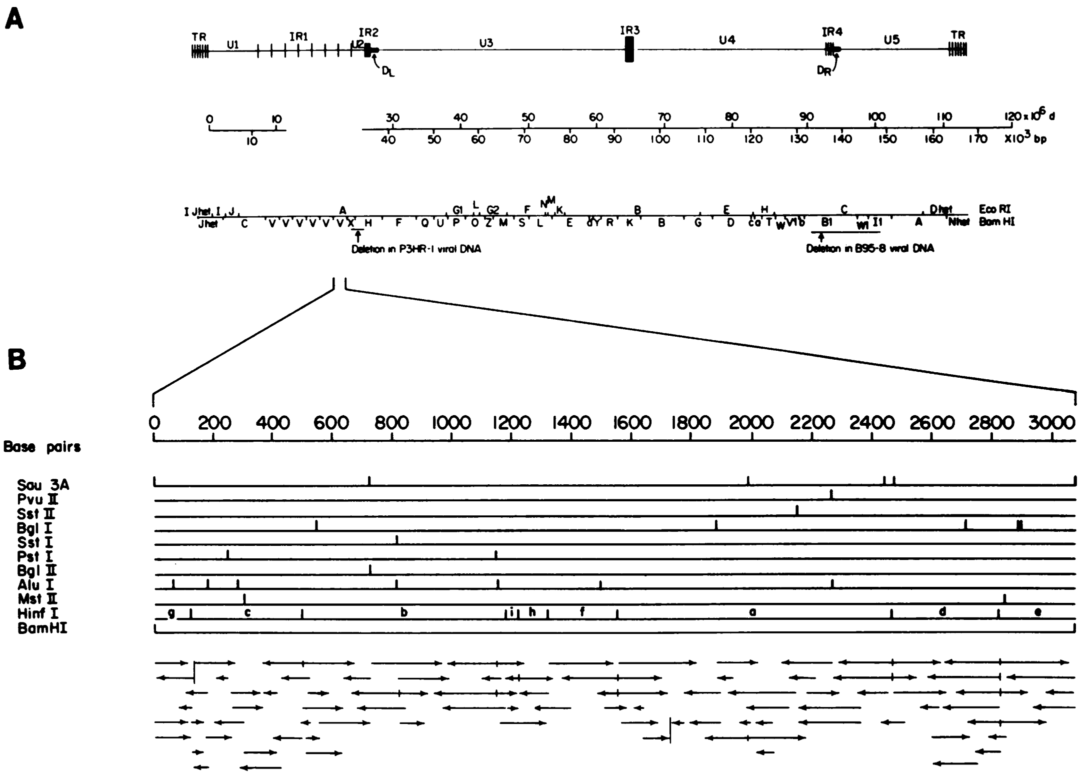

FIG. 1. (A) Overall structure of EBV DNA (6–8, 32) and (B) strategy for sequencing the *Bam*HI-V DNA fragment. DNA fragments were 3′ or 5′ end labeled after cleavage by various restriction enzymes. The length of the arrows represents the number of nucleotides determined from a set of sequencing reactions.

*Bam*HI fragments of EBV (B95-8) DNA cloned into the *Bam*HI site of pBR322, were grown in X1776 under P1 EK2 conditions (5). A cosmid-*Eco*RI-A clone of EBV (W91) DNA was grown in HB101 under P1 EK2 conditions (32). *Bam*HI-C is the juncture fragment between the last *Bam*HI site in U1 and the first site in IR1 (Fig. 1A). *Bam*HI-V is the fragment formed by cleavage at the *Bam*HI sites of adjacent tandem repeats of IR1 (Fig. 1A). EBV *Eco*RI-A contains part of U1 and U3 and all of IR1, U2, and IR2 (Fig. 1A).

Separation of restriction enzyme fragments. Restriction enzymes were obtained commercially (Bethesda Research Laboratory, New England Biolabs, or P-L Biochemicals, Inc.) and used under conditions specified by the manufacturers. Digests of cloned EBV DNA were separated on 5% polyacrylamide gels in a buffer consisting of 100 uM Tris-borate (pH 8.3) and 1 mM EDTA. DNA fragments purified from polyacrylamide gels were used for nucleotide determination, strand separation, or further enzyme digestion.

DNA sequence analysis. DNA fragments were either 5′ end labeled with bacterial alkaline phosphotase, [^{32}P]ATP and polynucleotide kinase or 3′ end labeled with ^{32}P-labeled cordycepin and terminal transferase (26). Enzymes were purchased from P-L Biochemicals, and radiochemicals were purchased from Amersham Corp. The end-labeled DNAs were purified, strand separated, or recut with another restriction enzyme and sequenced by chemical degradation (26). The nucleotide sequences were analyzed on an Amdahl computer, using previously derived programs (24, 31).

RESULTS

Nucleotide sequence of *Bam*HI-V DNA. Both strands of *Bam*HI-V from clone pDK14 were sequenced throughout, using 5′ or 3′ end-labeled DNA, strand separated or recut with a restriction enzyme, and the chemical degradation method. The sequencing strategy for the entire region is shown in Fig. 1B and involves the determination of over 100 nucleotide sequences. This large number was in part necessitated by the high guanine plus cytosine content of *Bam*HI-V (66.8%), which resulted in compression artifacts of some sequences which were particularly high in guanine and cytosine. Every nucleotide was determined from at least two sets of sequencing gels. Overlapping sequences were determined through every part of IR1 to avoid the possibility of omitting a small fragment. The

```
        10         20         30         40         50         60         70
GATCCCCCCA CCGGCCCTTC TCTCTGTCCC CCTGCTCCTC TCCAACCTTC GCTCCACCCT AGACCCCAGC
        80         90        100        110        120        130        140
TTCTGGCCTC CCCGGGTCCA CCAGGCCAGC CGGAGGGACC CCGGCAGCCC GGGCGAGTCG CCTTCCCTCT
       150        160        170        180        190        200        210
CCCCTGGCCT CTCCTTCCCG CCTCCCACCC GAGCCCCCTC AGCTTGCCTC CCCACCGGGT CCATCAGGCC
       220        230        240        250        260        270        280
GGCCGGAGGG ACCCCGGCGG CCCGGTGTCA GTCCCCCCTG CAGCCGCCCA GTCTCTGCCT CCAGGCAAGG
       290        300        310        320        330        340        350
GCGCCAGCTT TTCTCCCCCC AGCCTGAGGC CCAGTCTCCT GTGCACTGTC TGTAAAGTCC AGCCTCCCAC
       360        370        380        390        400        410        420
GCCCGTCCAC GGCTCCCGGG CCCAGCCTCG TCCACCCCTC CCCACGGTGG ACAGGCCCTC TGTCCACCCG
       430        440        450        460        470        480        490
GGCCATCCCC GCCCCCCTGT GTCCACCCCA GTCCCGTCCA GGGGGGACTT TATGTGACCC TTGGGCCTGG
       500        510        520        530        540        550        560
CTCCCCATAG ACTCCCATGT AAGCCTGCCT CGAGTAGGTG CCTCCAGAGC CCCTTTTGCC CCCCTGGCGG
       570        580        590        600        610        620        630
CCCAGCCCGA CCCCCGGGCG CCCCCAAACT TTGTCCAGAT GTCCAGGGGT CCCCGAGGGT GAGGCCCAGC
       640        650        660        670        680        690        700
CCCCTCCCGC CCCTGTCCAC TGCCCCGGTC CCCCCAGAAG CCCCCAAAAG TAGAGGCTCA GGCCATGCGC
       710        720        730        740        750        760        770
GCCCTGTCAC CAGGCCTGCC AAAGAGCCAG ATCTAAGGCC GGGAGAGGCA GCCCCAAAGC GGGTGCAGTA
       780        790        800        810        820        830        840
ACAGGTAATC TCTGGTAGTG ATTTGGACCC GAAATCTGAC ACTTTAGAGC TCTGGAGGAC TTTAAAACTC
       850        860        870        880        890        900        910
TAAAAATCAA AACTTTAGAG GCGAATGGGC GCCATTTTGT CCCCACGCGC GCATAATGGC GGACCTAGGC
       920        930        940        950        960        970        980
CTAAAACCCC CAGGAAGCGG GTCTATGGTT GGCTGCGCTG CTGCTATCTT TAGAGGGGAA AAGAGGAATA
       990       1000       1010       1020       1030       1040       1050
AGCCCCCAGA CAGGGGAGTG GGCTTGTTTG TGACTTCACC AAAGGTCAGG GCCCAAGGGG GTTCGCGTTG
      1060       1070       1080       1090       1100       1110       1120
CTAGGCCACC TTCTCAGTCC AGCGCGTTTA CGTAAGCCAG ACAGCAGCCA ATTGTCAGTT CTAGGGAGGG
      1130       1140       1150       1160       1170       1180       1190
GGACCACTGC CCCTGGTATA AAGTGGTCCT GCAGCTATTT CTGGTCGCAT CAGAGCGCCA GGAGTCCACA
      1200       1210       1220       1230       1240       1250       1260
CAAATGTAAG AGGGGGTCTT CTACCTCTCC CTAGCCCTCC GCCCCCTCCA AGGACTCGGG CCCAGTTTCT
      1270       1280       1290       1300       1310       1320       1330
AACTTTTCCC CTTCCCTCCC TCGTCTTGCC CTGCGCCCGG GGCCACCTTC ATCACCGTCG CTGACTCCGC
      1340       1350       1360       1370       1380       1390       1400
CATCCAAGCC TAGGGGAGAC CGAAGTGAAG GCCCTGGACC AACCCGGCCC GGGCCCCCCG GTATCGGGCC
      1410       1420       1430       1440       1450       1460       1470
AGAGGTAAGT GGACTTTAAT TTTTTCTGCT AAGCCCAACA CTCCACCACA CCCAGGCACA CACTACACAC
      1480       1490       1500       1510       1520       1530       1540
ACCCACCCGT CTCAGGGTCC CCTCGGACAG CTCCTAAGAA GGCACCGGTC GCCCAGTCCT ACCAGAGGGG
      1550       1560       1570       1580       1590       1600       1610
GCCAAGAACC CAGACGAGTC CGTAGAAGGG TCCTCGTCCA GCAAGAAGAG GAGGTGGTAA GCGGTTCACC
      1620       1630       1640       1650       1660       1670       1680
TTCAGGGGTA AGTAACCTGA CCTCTCCAGG GCTCACATAA AGGGAGGCTT AGTATACATG CTTCTTGCTT
      1690       1700       1710       1720       1730       1740       1750
TTCACAGGAA CCTGGGGGCT AGTCTGGGTG GGATTAGGCT GCCTCAAGTT GCATCAGCCA GGGCTTCATG
      1760       1770       1780       1790       1800       1810       1820
CCCTCCTCAG TTCCCTAGTC CCCGGGCTTC AGGCCCCCTC CGTCCCCGTC CTCCAGAGAC CCGGGCTTCA
      1830       1840       1850       1860       1870       1880       1890
GGCCCTGCCT CTCCTGTTAC CCTTTTAGAA CCACAGCCTG GACACATGTG CCAGACGCCT TGGCCTCTAA
      1900       1910       1920       1930       1940       1950       1960
GGCCCTCGGG TCCCCCTGGA CCCCGGCCTC AGCAACCCTG CTGCTCCCCT CCTGCCACCC CAGCCTCCCC
      1970       1980       1990       2000       2010       2020       2030
CCCTCCCCGT CCCCCTTCGC TCCTGATCCT CCCCCGGTCC CCAGTAGGGC CGCCTGCCCC CCTGCACCCA
      2040       2050       2060       2070       2080       2090       2100
GTACCTGCCC CTCTTGGCCA CGCACCCCGG GCCAGGCCAC CTTAGACCCG GCCAAGCCCC ATCCCTGAAG
      2110       2120       2130       2140       2150       2160       2170
ACCCAGCGGC CATTCTCTCT GGTAACGAGC AGAGAAGAAG TAGAGGCCCG CGGCCATTGG GCCCAGATTG
      2180       2190       2200       2210       2220       2230       2240
AGAGACCAGT CCAGGGGCCC GAGGTTGGAG CCAGCGGGCA CCCGAGGTCC CAGCACCCGG TCCCTCCGGG
      2250       2260       2270       2280       2290       2300       2310
GGGCAGAGAC AGGCAGGGCC CCCCGGCAGC TGGCCCCGAG GAGGCGCCCG GAGTGGGGCC GGTCGGCTGG
      2320       2330       2340       2350       2360       2370       2380
GCTGGCCGAG CCCGGGTCTG GGAGGTCTGG GGTGGCGAGC CTGCTGTCTC AGGAGGGGCC TGGCTCCGCC
      2390       2400       2410       2420       2430       2440       2450
GGGTGGCCCT GGGGTAAGTC TGGGAGGCAG AGGGTCGGCC TAGGCCCGGG GAAGTGGAGG GGGATCGCCC
      2460       2470       2480       2490       2500       2510       2520
GGGTCTCTGT TGGCAGAGTC CGGGCGATCC TCTGAGACCC TCCGGGCCCG GACTGTCGCC CTCAGCCCCC
      2530       2540       2550       2560       2570       2580       2590
CAGACAGACC CCAGGGTCTC CAGGCAGGGT CCGGCATCTT CAGGGGCAGC AGGCTCACCA CCACAGGCCC
      2600       2610       2620       2630       2640       2650       2660
CCCAGACCCG GGTCTCGGCC AGCCGAGCCG ACCGGCCCGC GCCTGGCGCC TCCTCGGGGC CAGCCGCCGG
      2670       2680       2690       2700       2710       2720       2730
GGTTGGTTCT GCCCCTCTCT CTGTCCTTCA GAGGAACCAG GGACCTCGGG CACCCCAG(AG) CCCCTCGGGC
      2740       2750       2760       2770       2780       2790       2800
CCGCCTCCAG GCGCCCTCCT GGTCTCCGCT CCCCTCTGAG CCCCGTTAAA CCCAAAGAAT GTCTGAGGGG
      2810       2820       2830       2840       2850       2860       2870
AGCCACCCTC GGGGCCCAGG CCCCAGAGTC CAGAGGTCAG GGGCACCTCA GGGTGCCTCC CCGGGTCCCA
      2880       2890       2900       2910       2920       2930       2940
GGCCAGCCGG AGGGACCCCG GCAGCCCGGG CGGCCCCAGA GGCCGGTTCC TCGCCCCTTC CCCGGGCTTC
      2950       2960       2970       2980       2990       3000       3010
AGAGCCCAGG ATGTCCCCCA GAAGGGACCC TAGGCGTCCC CTCTCCTCCC CTCCAGGCCC GAGCCTCTCC
      3020       3030       3040       3050       3060       3070
CTCGCGGAGA GGGGCCTCTT TGGGCCCTCA AGTCCAGCCC CACCGAGACC CGAGTGGCCC G
```

FIG. 2. Nucleotide sequence of IR1 from *Bam*HI-V of pDK14. The nucleotide sequence is shown 5′ to 3′, from the unique *Bam*HI site in one copy of IR1 to the unique *Bam*HI site in the rightward next copy (orientation as shown in Fig. 1). Potential transcriptional signals CCAAT, TATAAA, and TATA are indicated by boxes. The sequence homologous to the papovavirus ori is underlined. The order of the dinucleotide (AG) is not certain.

entire *Bam*HI-V sequence is 3,071 bp. The sequence is described in Fig. 2, 5′ to 3′ in the map orientation indicated in Fig. 1 (Fig. 2). All coordinates in this manuscript are relative to the coordinates shown in Fig. 2. Previously, the size of *Bam*HI-V was inaccurately estimated to be 3,360 bp because of erroneous determination of the size of the largest *Hin*fI fragments (3).

The identical 35-nucleotide (CCAGGCCAGCCGGAGGGACCCCGGCAGCCCGGGCG) direct repeat at 91 and 2,868 and a similar 35-base sequence at 194 (3) were previously identified. There are three other long direct repeats, a 16-bp repeat and two 14-bp repeats. The 16-bp repeat CCCGGGCTTCAGGCCC has its 5′ end at 1,771 and 1,810. One 14-bp repeat GCCTCCCCGGGTCC has its 5′ end at 76 and 2,855. The other 14-bp repeat TCCCCGGGCTTCAG is at 1,769 and 2,929. The 14-bp repeat at 76 is 1 bp before the 35-base repeat at 91, and its counterpart at 2,855 is 2 bp before the 35-base repeat at 2,868. The 14- and 35-bp perfect repeats could therefore be considered as imperfect direct repeats of 51 bases which contribute to the hybridization between the ends of *Bam*HI-V which was previously reported (3). The direct repeats are shown schematically in Fig. 3.

There are 12 perfect dyad symmetries of more than 10 bp that are separated by various numbers of nucleotides (Fig. 3). Six of these dyad symmetries are between bases 2,211 and 2,710. The longest dyad symmetry is 20 bases, (GCTGGCCCCGAGGAGGCGCC at 2,269 to 2,288 and CGACCGGGCTCCTCCGCGG, at 2,654 to 2,635) and is part of the 6-dyad symmetry cluster.

Comparison of *Bam*HI-V clones. pDK14 and pDK51 are independently isolated clones of EBV B95-8 *Bam*HI-V in pBR322 (5). Double digestion of both DNAs with *Bam*HI-*Hin*fI gives the same DNA fragments when analyzed on 1% agarose or 5% polyacrylamide gels (Fig. 4). The nucleotide sequence of both strands of the pDK51 *Bam*HI-V and *Hin*fI-h and *Hin*fI-i fragments were determined and are identical to the same fragments of pDK14. *Bam*HI-V was also isolated from a clone of EBV W91 *Eco*RI-A fragment in MUA3 (32). The nucleotide sequence of 130 nucleotide from each 5′ *Bam*HI labeled end was identical to the sequence of pDK14 except for a G at 11 and a C at 3,041 (data not shown).

Juncture of U1-IR1. *Bam*HI-C contains the U1-IR1 juncture and part of IR1 before the first

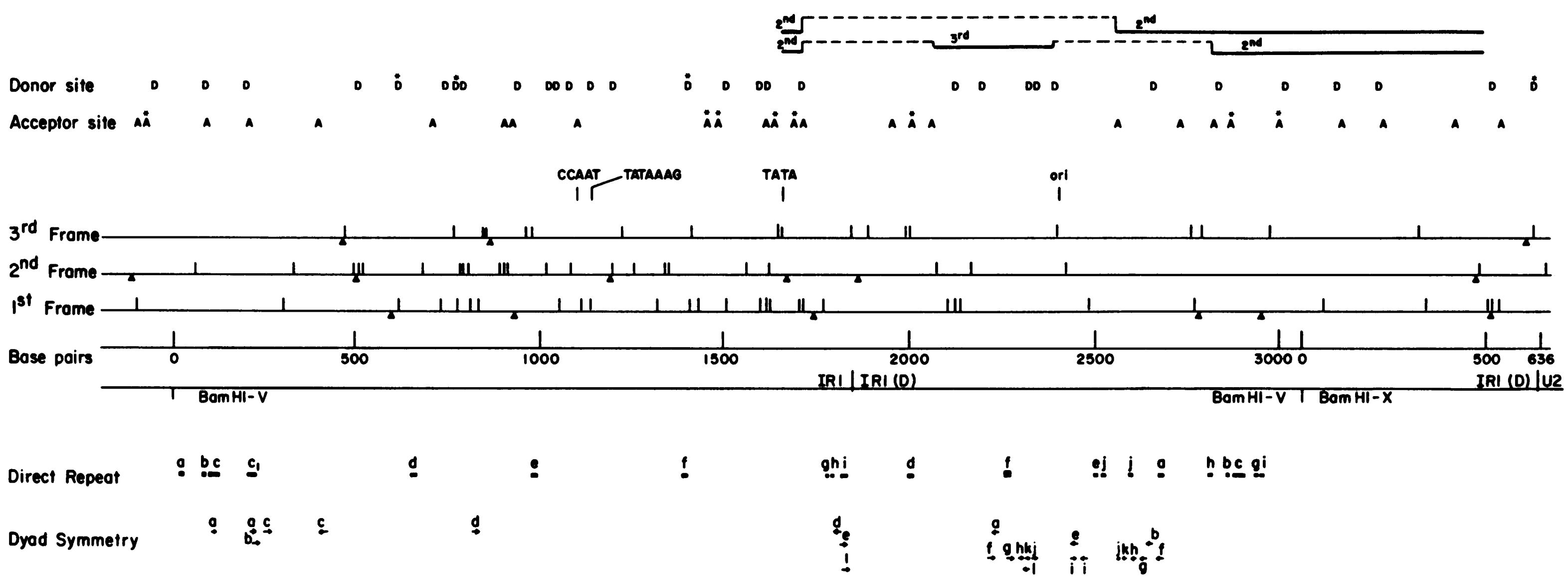

FIG. 3. Schematic representation of IR1 and the adjacent few nucleotides of U2 indicating splice sites, SV40 *ori* homologous sequence, potential transcriptional signals, initiating codons, stop codons, direct repeats greater than 10 bp, and dyad symmetries greater than 10 bp. The top panel shows two potential transcripts. Introns and exons are presented by (---) and (——), respectively. (D, Ḋ) and (A, Ȧ) are donor and acceptor sites at splice junctions as explained in the text (21). D indicates 6 bp and Ḋ indicates more than 6 bp identity to a known donor sequence. Ȧ indicates a sequence which conforms to the consensus acceptor sequences, and A indicates a less likely acceptor sequence which is not AT rich or contains the dinucleotide AG at the boundary. The three reading frames are arbitrarily assigned. Termination codons (TAA, TAG, and TGA) are represented by (⊥), and AUG triplets are indicated by (▲). Homologous regions (direct repeats and dyad symmetries) are indicated by matching letters.

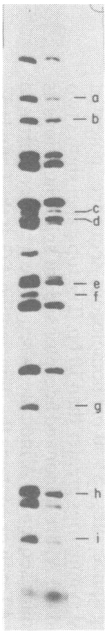

FIG. 4. Comparison of pDK14 and pDK51 clones. DNAs from both clones were cleaved with *Hin*fI and *Bam*HI and electrophoresed in a 5% polyacrylamide gel in a 500 mM Tris-borate (pH 8.3) and 10 mM EDTA buffer. The letters indicated *Hin*fI fragments of the *Bam*HI-V insert. The *Hin*fI-h and *Hin*fI-i fragments of both clones were sequenced.

*Bam*HI site in IR1. Enzymes which cut *Bam*HI-C between the beginning of IR1 and the first *Bam*HI site should yield the same size fragment from *Bam*HI-C and *Bam*HI-V. Previous data indicated that *Bam*HI-C contains less than 2,200 bp of IR1 (3). To more precisely map the transition, *Bam*HI-C and *Bam*HI-V DNA were purified, 5′ end labeled, and digested with *Sst*I. *Sst*I does not cleave within the IR1 portion of *Bam*HI-C, but does cleave *Bam*HI-C in U1 and *Bam*HI-V in the part of IR1 missing from *Bam*HI-C (3). The end-labeled fragments of *Bam*HI-V and *Bam*HI-C were isolated and incubated with *Mob*II (Fig. 5A) or with *Ava*II (Fig. 5B) for short intervals so as to achieve limited digestion. The extent of conservation of IR1 in *Bam*HI-C was evaluated from the similarity in size of the partial digestion products of *Bam*HI-C and *Bam*HI-V, both of which were labeled at their right ends. The data suggest that *Bam*HI-C and *Bam*HI-V share at least the same restriction endonuclease sites for 1,172 bp (*Ava*II site, Fig. 5B) from their respective rightward *Bam*HI restriction site.

Sequence of the U1-IR1 transition. *Bgl*I cuts both *Bam*HI-V and *Bam*HI-C 9 bp to the left of the *Ava*II site which is 1,172 bp before the *Bam*HI site (Fig. 5C). The *Bgl*I fragment of *Bam*HI-C, which is 1,181 bp to the left of the *Bam*HI site, was 3′ end labeled, purified, and sequenced. Except for a base difference, indicated in Fig. 6, the IR1 nucleotide sequence of *Bam*HI-C and *Bam*HI-V are identical for 28 bases, including the left portion of the *Bgl*I recognition site (GCCNNNNNGGC). Thus, the transition from U1 to IR1 occurs 1,214 nucleotides 5′ to the first *Bam*HI site in IR1 (Fig. 7).

DISCUSSION

The data summarized in Fig. 7 illustrate that the transition from U1 to IR1 occurs in *Bam*HI-C, 1,214 bp before the first *Bam*HI-V fragment, and the transition from IR1 to U2 occurs in *Bam*HI-X, 636 bp after the last *Bam*HI-V fragment. Therefore, relative to the start of the first copy of IR1 at the juncture with U1, the last IR1 is 1,221 bp short of being complete. There is no apparent structural specificity for the deletion in the last (or first copy) of IR. There are no direct repeats or dyad symmetries within IR1 at either juncture. There is no extensive homology between the adjacent U1 and U2 sequences or between these sequences and the sequences within IR1, such as frequently occurs at junc-

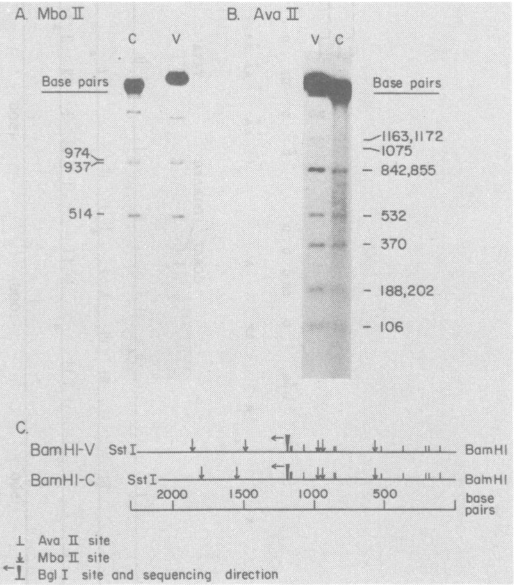

FIG. 5. Mapping the U1-IR1 juncture by comparison of partial digestion products of *Bam*HI-C and *Bam*HI-V (37). The right ends of *Bam*HI-V and *Bam*HI-C were obtained after 5′ end labeling and digestion with *Sst*I. The *Bam*Hi end-labeled fragments were subjected to partial digestion by (A) *Mbo*II or (B) *Ava*II. The size of fragments was determined by their electrophoretic mobility and the known nucleotide sequence of *Bam*HI-V. (C) Summary of partial digestion results and the strategy employed for the determination of U1-IR1 juncture.

```
A. Bam HI C

     40         30         20         10    *
TCT  CACAAAAACA TTCTGGACAC ATGTGCCAGA CGCCTGGGCC

B. Bam HI V

     40         30         20         10    *
TTT  TAGAACCACA GCCTGGACAC ATGTGCCAGA CGCCTTGGCC
```

FIG. 6. Nucleotide sequence to the left of the *Bgl*I sites in *Bam*HI-C and *Bam*HI-V. The DNAs were 3′ end labeled at the *Bgl*I site and sequenced by the chemical degradation method. The leftmost portion (GCC) of the *Bgl*I site (GCCNNNNNGGC) is shown. Nucleotides common to *Bam*HI-C and *Bam*HI-V are underlined.

tures of repeat and unique DNAs and at sites of insertion of transposable elements (15; J. Shapiro and B. Cordell, Biol. Cell, in press).

Single-base differences might be anticipated among some or all of the IR1 copies in EBV DNAs, since low-level variation in restriction endonuclease sites has been observed to occur in other parts of EBV DNA (6, 8, 17, 32). Variation among the copies of a repeat sequence of a single isolate would be expected to be unstable and should be eliminated or spread through all copies (36). No differences were found in the IR1 portion of *Bam*HI-V and *Bam*HI-X which was previously sequenced (3) or among the *Bam*HI-V *Hin*fI-h and *Hin*fI-i fragments of two different clones of B95-8 *Bam*HI-V (pDK14 and pDK51). The IR1 portion of *Bam*HI-C and *Bam*HI-V were found to differ by a single base (Fig. 2), and the terminal 260 nucleotides of *Bam*HI-V of W91 DNA has two single-base differences from B95-8 *Bam*HI-V. The differences observed could reflect base differences between the first and repeat copies of IR1 (*Bam*HI-C and *Bam*HI-V difference) and differences among isolates (W91 and B95-8 difference), or could reflect changes introduced during cloning into pBR322 and growth in *Escherichia coli*.

General schemes for eucaryotic transcriptional signals have been outlined (1, 4). A sequence, CCAAT-39 bp-TATAAA, similar to the prototype occurs at 1,098 bp (Fig. 2 and 3). A TATA sequence occurs at 1,662 bp, but is not preceded by CCAAT and is less likely to be part of a functional promoter. There is no polyadenylation signal, −AATAAA−, in IR1 (1). The first potential translation initiation codon (AUG) 3′ to the TATAAA is at 1,193 bp. A termination codon (TAA) occurs immediately after the AUG. Therefore, this AUG is unlikely to be used. The first AUG after either TATA sequence which is not immediately blocked by a termination codon is located at 1,668 bp. The only long open reading frame is in the second frame and extends from 2,421 to 474 bp (Fig. 2 and 3).

Potential donor and acceptor sites (Fig. 3) were designated based on the splice junction catalog compiled by Mount (27). The donor sequences (D and Ḋ) are identical to at least one known junction sequence and the acceptor sequences (A and Ȧ) are similar to the consensus (T/C)$_n$ N(T/C)AGG sequence (Fig. 3). Two mRNAs could be encoded by IR1 utilizing the putative promotor at 1,098 bp, the AUG at 1,668 bp, the long open reading frame and splice sequences to circumvent termination (Fig. 3). One of the RNAs would use only the second frame. The other would use parts of both the second and the third frames. However, to complete a 3-kb RNA, at least one additional splice is required into the U2 region. Recent studies of the RNA encoded by IR1 and U2 in latently infected cells indicate that it is transcribed from left to right in the orientation shown and is spliced (van Santen and Kieff, manuscript in preparation).

The origin of EBV DNA replication is not known. In productive infection, viral DNA is believed to be replicated by a virus-specified polymerase, since viral DNA synthesis is inhibited by drugs which have little effect on cell DNA synthesis (38, 39). In latently infected cells, EBV DNA appears to be replicated by cellular DNA polymerase, since viral DNA synthesis is resistant to usual inhibitors (38, 39) and viral DNA is synthesized during the normal S phase (14). Since, in latent infection, viral DNA is probably replicated by cellular polymerase, the origin of EBV DNA replication might be expected to be similar to origins of cell DNA synthesis. A 19-nucleotide sequence, GGAGGCAGAGGGTCGGCCT, at 2,402 to 2,421 in IR1 is similar to the sequences which occur at the origins for SV40 (13, 33, 35) and BK virus (41) DNA synthesis (Fig. 8). The papovavirus *ori* sequences are homologous to sequences in the "*Alu* family" of interspersed repeated cell DNA sequences, which may be origins of cell DNA synthesis (15, 21, 28; M. Singer, Int. Rev. Cytol., in press).

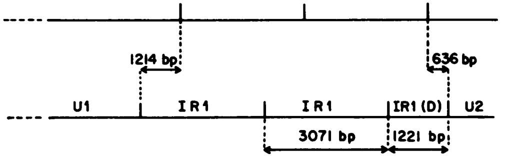

FIG. 7. Summary of the organizational features around IR1. *Bam*HI-C consists of the right end of U1 and the left portion of the first IR1. *Bam*HI-V is the fragment generated by cleaving adjacent IR1 repeats with *Bam*HI. *Bam*HI-X consists of the end portion of IR and the beginning of U2 (3). IR1 (D) is the last and incomplete copy of IR1; its length is 1,850 bp, not 1,221 bp as indicated in the Figure.

```
SV40   5' CTACTTCTGGAATAGCTC|AGAGGCCGAGG   CGG|CCTC 3'
BK     5' CTACTTGAGAGAAAGGGT|GGAGGCAGAGG   CGG|CCTC 3'
EBV          5' TGGGGTAAGTCTG|GGAGGCAGAGG   CGG|CCTA 3'
                             |          \  /   |
                             |           G-T   |
```

FIG. 8. Homology of EBV IR1 nucleotides 2403 to 2421 to SV40 (13, 33, 35) and BK (41) virus *ori* regions.

Several of the dyad symmetries in EBV DNA cluster in the 2,211 to 2,700 region, which includes the *ori* homologous site (Fig. 3). The possible secondary single-strand structures around this region were evaluated. Perfect dyad symmetries greater than 10 bp in this region are shown in Fig. 9A. Four of the dyad symmetries would form a nearly perfectly matched stem, one side of which extends from 2,269 to 2,332 bp and the other which extends from 2,591 to 2,654 bp. If imperfect dyad symmetries are considered, the structure becomes a giant hairpin extending from 2,205 to 2,458 bp on one side and 2,463 to 2,714 bp on the other (Fig. 9B). At the extremity of the dyad symmetry is a 9-bp direct repeat with 5′ ends at 2,205 and 2,707 bp. Part of direct repeat (D1, Fig. 9) is also part of the final

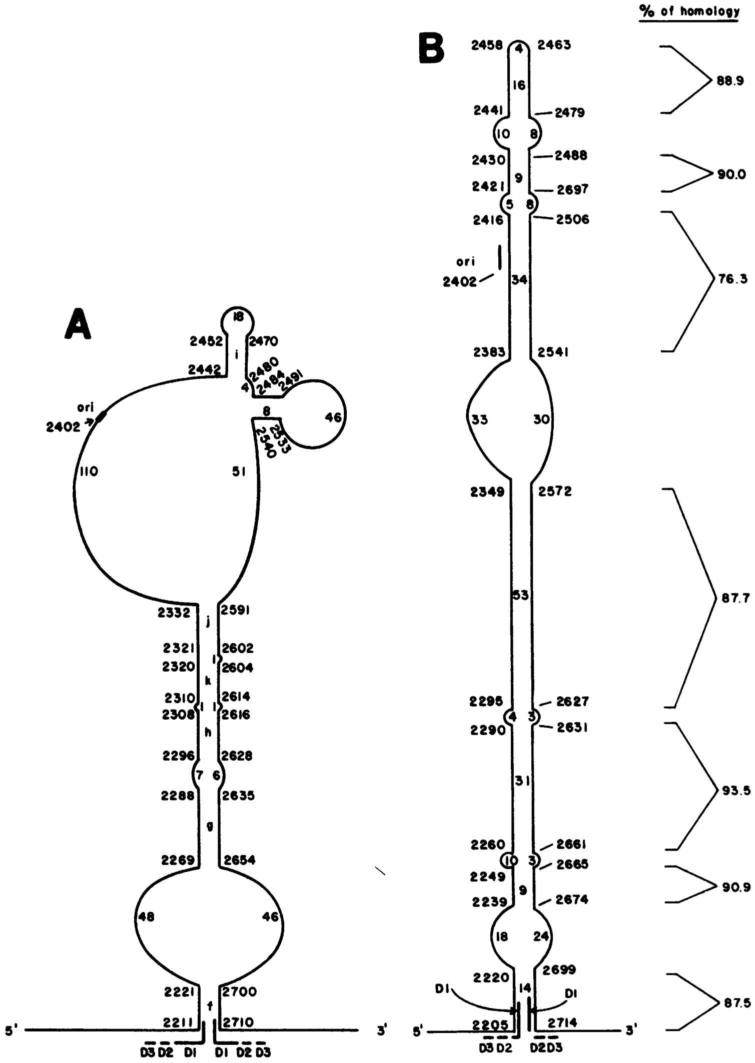

FIG. 9. Secondary structure around the ori homologous region anticipated from dyad symmetries. (A) Perfect dyad symmetries. The number of base pairs in each symmetrical region is indicated as follows: f, 11; g, 20; h, 13; k, 11; i, 11; and j, 12. (B) Secondary structure of this region when imperfect dyad symmetries are considered. D1 indicates the 9-bp direct repeat flanking the entire region, and D2 and D3 are each 3-bp direct repeats immediately adjacent to D1.

11-bp perfect dyad symmetry which flanks the hairpin (Fig. 9). Adjacent to the 9-bp direct repeat are two nucleotide triplets which are also directly repeated at both ends of the dyad symmetry (Fig. 9).

As suggested above, the IR1 dyad symmetry is similar to the *Alu* family of interspersed repeated DNA sequences in its homology to papovavirus origins of DNA replication (15, 21, 28, 33, 41; M. Singer, in press). The *Alu* family sequence is approximately 300 bp, whereas each side of the IR1 dyad symmetry is 240 bp. The *Alu* family sequence usually occurs as interspersed direct repeats or as inverted repeats separated by intervening unique sequences, but also occurs as head-to-head palindromes (600 bp overall [9] compared with the 500-bp IR1 dyad symmetry). The *Alu* family sequence is partly composed of two long direct repeats. No similar direct repeat has been discerned within each side of the IR1 dyad symmetry. The IR1 dyad symmetry also resembles transposable elements by having inverted repeat sequences at the ends (Shapiro and Cordell, in press). The dinucleotides 5′ TG . . . . . . CA 3′ are frequently found at the ends of eucaryotic transposable elements and are not a feature of the IR1 dyad or of the *Alu* family (Shapiro and Cordell, in press).

The two sides of the dyad symmetry are likely to have originated from a common sequence since the palindrome is highly conserved. Since IR1 has a similar size in the related herpesvirus papio and herpesvirus pan DNAs (18–20, 25), the evolution of the palindrome must have preceded the divergence of the old-world primate, EBV-related viruses. The similarities of the dyad symmetry to *Alu* family sequences suggests the possibility that this part of the IR1 may have evolved from a cellular sequence.

The high degree of conservation of the EBV IR1 palindrome at the stem 63-bp region (Fig. 9A) suggests that either there is ongoing base pairing between the components of the palindrome or the two 63-bp regions are rigidly fixed in nucleotide sequence, such as might occur if the sequence is a recognition site for a regulatory protein. We favor the latter possibility. The dyad symmetry cluster could give rise to localized cruciforms (29) near the central axis (Fig. 9B) but to melt out the entire region and form the structures shown in Fig. 9 is thermodynamically impossible.

ACKNOWLEDGMENTS

We thank Tim Dambaugh for clones of EBV DNA, Chris Beisel and Susan Fennewald for help with computer analysis of the nucleotide sequence, Larvan Chamnankit for technical assistance in DNA sequencing, and Jim Shapiro, Tom Petes, and Nick Cozzarelli for helpful discussions.

This research was supported by Public Health Service grants CA 19264 and CA 17281 from the National Cancer Institute and by American Cancer Society grant MV 32G. E.K. is a Faculty Research Awardee of the American Cancer Society.

ADDENDUM IN PROOF

The large dyad symmetry described in IR1 (Fig. 9) would be expected to have a profound effect on the secondary structure of the RNA transcribed from this region. Shorter dyad symmetries which are also guanine- and cytosine-rich act as terminators of transcription of SV40 late RNA and may be involved in the attenuation of transcription or in the modulation of translation or RNA processing (N. Hay, H. Skolnick-David, and Y. Aloni, Cell **29**:183–193, 1982).

LITERATURE CITED

1. **Benoist, C., K. O'Hare, R. Breathmach, and P. Chambon.** 1980. The ovalbumin gene: sequence of putative control regions. Nucleic Acids Res. **8**:127–142.
2. **Bornkamm, G. W., H. Delius, U. Zimber, 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.
3. **Cheung, A., and E. Kieff.** 1981. Epstein-Barr virus DNA. X. Direct repeat within the internal direct repeat of Epstein-Barr virus DNA. J. Virol. **40**:501–507.
4. **Corden, J., A. Wasylyk, A. Buckwalder, P. Sassone-Corsi, C. Kedinger, and P. Chambon.** 1980. Promoter sequences of eukaryotic protein-coding genes. Science **209**:1406–1414.
5. **Dambaugh, T., C. Beisel, M. Hummel, W. King, S. Fennewald, A. Cheung, M. Heller, N. Raab-Traub, and E. Kieff.** 1980. Epstein-Barr virus DNA. VII. Molecular cloning and detailed mapping of EBV (B95-8) DNA. Proc. Natl. Acad. Sci. U.S.A. **77**:2999–3003.
6. **Dambaugh, T., M. Heller, N. Raab-Traub, W. King, A. Cheung, C. Beisel, M. Hummel, V. van Santen, S. Fennewald, and E. Kieff.** 1980. DNAs of Epstein-Barr virus and herpes virus papio, p. 85–90. *In* A. Nahmias, W. Dondle, and R. Schinazi (ed.), The human herpes virus. Elsevier, New York.
7. **Dambaugh, T., F. Nkrumah, R. J. Biggar, and E. Kieff.** 1979. Epstein-Barr virus RNA in Burkitt tumor tissue. Cell **16**:313–322.
8. **Dambaugh, T., N. Raab-Traub, M. Heller, C. Beisel, M. Hummel, A. Cheung, S. Fennewald, W. King, and E. Kieff.** 1980. Variations among isolates of Epstein-Barr virus. Ann. N.Y. Acad. Sci. **602**:711–719.
9. **Deininger, P., and C. Schmid.** 1976. An electron microscope study of the DNA sequence organization of the human genome. J. Mol. Biol. **106**:773–790.
10. **Given, D., and E. Kieff.** 1978. DNA of Epstein-Barr virus. IV. Linkage map for restriction enzyme fragments of the B95-8 and W91 strains of EBV. J. Virol. **28**:524–542.
11. **Given, D., and E. Kieff.** 1979. DNA of Epstein-Barr virus. VI. Mapping of the internal tandem reiteration. J. Virol. **31**:315–324.
12. **Given, D., D. Yee, K. Griem, and E. Kieff.** 1979. DNA of Epstein-Barr virus. V. Direct repeats at the ends of EBV DNA. J. Virol. **30**:852–862.
13. **Gluzman, Y., J. Sambrook, and R. Frisque.** 1980. Expression of early genes of origin-defective mutants of simian virus 40. Proc. Natl. Acad. Sci. U.S.A. **77**:3898–3902.
14. **Hampar, B., A. Tanaka, M. Nonoyama, and J. Derge.** 1974. Replication of the resident repressed EBV genome during early S phase (S-1 period) of non-producer Raji cells. Proc. Natl. Acad. Sci. U.S.A. **71**:631–633.
15. **Haynes, S. R., T. P. Toomey, L. Leinwand, and W. R. Jelinek.** 1981. The Chinese hamster Alu-equivalent sequence: a conserved, highly repetitious, interspersed DNA sequence in mammals has a structure suggestive of a transposable element. Mol. Cell. Biol. **1**:573–583.

16. **Hayward, S., L. Nogee, and G. Hayward.** 1980. Organization of repeated regions within the Epstein-Barr virus DNA molecule. J. Virol. **33**:507–521.
17. **Heller, M., T. Dambaugh, and E. Kieff.** 1981. Epstein-Barr virus DNA. IX. Variation among viral DNAs. J. Virol. **38**:632–648.
18. **Heller, M., P. Gerber, and E. Kieff.** 1981. Herpes virus papio DNA is similar in organization to Epstein-Barr virus DNA. J. Virol. **37**:698–709.
19. **Heller, M., P. Gerber, and E. Kieff.** 1982. DNA of herpesvirus pan, a third member of the Epstein-Barr virus–herpesvirus papio group. J. Virol. **41**:931–939.
20. **Heller, M., and E. Kieff.** 1981. Colinearity between the DNAs of Epstein-Barr virus and herpesvirus papio. J. Virol. **37**:821–826.

20a. **Heller, M., V. van Santen, and E. Kieff.** 1982. Simple repeat sequence in Epstein-Barr virus DNA is transcribed in latent and productive infections. J. Virol. **44**:311–320.

21. **Jelinek, W. R., T. P. Toomey, L. Leinwand, C. Duncan, P. A. Biro, P. V. Choudary, S. Weissman, C. M. Rubin, C. M. Houck, P. L. Deininger, and C. W. Schmid.** 1980. Ubiquitous, interspersed repeated sequences in mammalian genomes. Proc. Natl. Acad. Sci. U.S.A. **77**:1398–1402.
22. **King, W., A. L. T. Powell, N. Raab-Traub, M. Hawke, and E. Kieff.** 1980. Epstein-Barr virus RNA. V. Viral RNA in a restringently infected, growth transformed cell line. J. Virol. **36**:506–518.
23. **King, W., V. van Santen, and E. Kieff.** 1981. Epstein-Barr virus RNA. VI. Viral RNA in restringently and abortively infected Raji cells. J. Virol. **38**:649–660.
24. **Korn, L., C. Queen, and M. Wegman.** 1977. Computer analysis of nucleic acid regulatory sequences. Proc. Natl. Acad. Sci. U.S.A. **74**:4401–4405.
25. **Lee, Y. S., M. Nonoyama, and H. Rabin.** 1981. Colinear relationships of herpesvirus papio DNA to Epstein-Barr virus DNA. Virology **110**:248–252.
26. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65**:499–560.
27. **Mount, S.** 1982. A catalogue of splice junction sequences. Nucleic Acids Res. **10**:459–472.
28. **Pan, J., J. T. Elder, C. H. Duncan, and S. M. Weissman.** 1981. Structural analysis of interspersed repetitive polymerase III transcription units in human DNA. Nucleic Acids Res. **9**:1151–1170.
29. **Panayotatos, N., and R. Wells.** 1981. Cruciform structures in supercoiled DNA. Nature (London) **289**:466–470.
30. **Powell, A. L. T., W. King, and E. Kieff.** 1979. Epstein-Barr virus specific RNA. III. Mapping of the DNA encoding viral specific RNA in restringently infected cells. J. Virol. **29**:261–274.
31. **Queen, C., and L. Korn.** 1980. Computer analysis of nucleic acid and proteins methods. Methods Enzymol. **65**:595–609.
32. **Raab-Traub, N., T. Dambaugh, and E. Kieff.** 1980. DNA of Epstein-Barr virus. VIII. Analysis and molecular epidemiology of the "additional" DNA of two Burkitt tumor isolates of EBV. Cell **22**:257–267.
33. **Reddy, V., B. Thimmappaya, R. Dhar, K. Subramanian, S. Zain, J. Pan, M. Celma, and S. Weissman.** 1978. The genome of simian virus 40. Science **200**:494–502.
34. **Rymo, L., and S. Forsblum.** 1978. Cleavage of Epstein-Barr virus DNA by restriction endonucleases EcoRI, HindIII and BamI. Nucleic Acids Res. **5**:1387–1402.
35. **Shortle, D., and D. Nathans.** 1979. Regulatory mutants of simian virus 40 constructed mutants with base substitutions at the origin of DNA replication. J. Mol. Biol. **131**:801–817.
36. **Smith, G. P.** 1976. Evolution of repeated DNA sequences by unequal crossover. Science **191**:528–535.
37. **Smith, H. O., and M. Birnstiel.** 1976. A simple method for DNA restriction site mapping. Nucleic Acids Res. **3**:2387–2398.
38. **Summers, W., and G. Klein.** 1976. Inhibition of Epstein-Barr virus DNA synthesis and late gene expression by phosphonoacetic acid. J. Virol. **18**:151–155.
39. **Thorley-Lawson, D., and J. L. Strominger.** 1976. Transformation of human lymphocytes by Epstein-Barr virus is inhibited by phosphonoacetic acid. Nature (London) **263**:332–334.
40. **van Santen, V., A. Cheung, and E. Kieff.** 1981. Epstein-Barr virus RNA. VII. Size and direction of transcription of virus specified cytoplasmic RNA in a cell line transformed by EBV. Proc. Natl. Acad. Sci. U.S.A. **78**:1930–1934.
41. **Yang, R., and R. Wu.** 1979. BK virus DNA: complete nucleotide sequence of human tumor virus. Science **206**:456–462.