JOURNAL OF VIROLOGY, July 1987, p. 2063–2070
0022-538X/87/072063-08$02.00/0

Vol. 61, No. 7

The 160,000-M_r Virion Protein Encoded at the Right End of the Herpesvirus Saimiri Genome Is Homologous to the 140,000-M_r Membrane Antigen Encoded at the Left End of the Epstein-Barr Virus Genome

KEITH R. CAMERON,[1] THOMAS STAMMINGER,[2] MOLLY CRAXTON,[1] WALTER BODEMER,[2] ROBERT W. HONESS,[1]* AND BERNHARD FLECKENSTEIN[2]

Division of Virology, National Institute for Medical Research, Mill Hill, London NW7 1AA, United Kingdom,[1] and Institut für Klinische Virologie, Universität Erlangen-Nürnberg, Erlangen, Federal Republic of Germany[2]

Received 3 December 1986/Accepted 14 March 1987

The sequence of 4.4 kilobase pairs (kbp) from the conventional right terminus of the A+T-rich light-DNA (L-DNA) sequences of the herpesvirus saimiri (HVS) genome contains a leftward-directed open reading frame (ORF) for a 1,299-residue protein. The molecular weight predicted for the protein (143,000) is in good agreement with the estimates of 150,000 to 160,000 for the major nonglycosylated polypeptide of the virion tegument (the 160K polypeptide), previously shown to be encoded by this region of the genome. The first initiation codon of the ORF is only 250 nucleotides from the junction of the L-DNA component with the G+C-rich terminal reiterations (i.e., heavy or H-DNA) of the genome. An unusually A+T-rich sequence (43 of 45 nucleotides are A or T, relative to a mean composition of 40% G+C for the ORF) occurs some 75 bp 5′ to this initiation codon, and the first adenylation signal (AATAAA) on this DNA strand occurs 18 bp 3′ to the termination codon. The amino acid sequence predicted for the 160K protein of HVS is homologous over most of its length to the 1,318-residue protein encoded by the leftmost major ORF of the G+C-rich genome of Epstein-Barr virus (BNRF1, the 140K nonglycosylated membrane antigen). No homology to either of these proteins is evident among the products predicted from the complete sequence of the alpha herpesvirus varicella-zoster virus. Thus gamma herpesviruses with coding sequences which differ in mean nucleotide composition by some 20% G+C have homologous proteins encoded at similar positions with respect to genome termini, with the right end of HVS being homologous to the left end of Epstein-Barr virus.

Herpesvirus genomes are all double-stranded DNA molecules of >100 kilobasepairs (kbp), but the mean nucleotide composition of genomes from different members of the group ranges from 36 to 75% G+C. Herpesviruses are also diverse in their biological properties, and a subdivision into alpha, beta, and gamma herpesvirus subgroups based upon general differences in these biological properties is proving useful and is supported by more quantitative measures of relatedness. Thus, the alpha herpesviruses are neurotropic viruses with either G+C-rich (α_1 subgroup; e.g., herpes simplex viruses, 66 to 68% G+C) or A+T-rich (α_2 subgroup; e.g., varicella-zoster virus, 46% G+C) coding sequences. The beta herpesviruses are synonymous with the salivary gland inclusion agents, or cytomegaloviruses (55 to 60% G+C), and the gamma herpesviruses include the G+C-rich lymphotropic herpesviruses of humans and old-world monkeys (e.g., Epstein-Barr virus [EBV], 60% G+C; γ_1 subgroup) and the lymphotropic herpesviruses of new-world monkeys and lower vertebrates, which have A+T-rich coding sequences (e.g., herpesvirus saimiri [HVS], 36% G+C; γ_2 subgroup) (20, 23, 43).

We have undertaken an analysis of the prototype virus of the distinctive γ_2 subgroup of lymphotropic herpesviruses, HVS. Although many general features of the biological properties of this virus are similar to those of EBV, the gross composition and structure of the genomes of these viruses differ markedly (1, 16, 26). The EBV genome is G+C-rich (60% G+C) and is characterized by large- and small-scale internal redundancy, notably in the form of multiple (more than nine) tandem reiterations of a >3-kbp sequence, the *Bam*HI-W repeats (1, 26). The herpesvirus saimiri genome has an A+T rich coding sequence (112 kbp; 36% G+C; i.e., L-DNA) with no large-scale internal redundancy, but with >30 reiterations of a 1.44-kbp sequence of 71% G+C at the termini of the genome (H-DNA; 3, 16, 27). However, studies of the gene products of HVS have revealed a marked similarity between many of the structural and nonstructural proteins of this virus and the proteins of EBV (41). For example, among their structural proteins, both viruses encode a nonglycosylated, or poorly glycosylated, protein of molecular weight 140,000 to 160,000, which is one of the major components which surrounds the capsid but is beneath the lipid envelope of their respective virions (i.e., the tegument [13, 14, 25, 39, 40]). In virions of EBV, this protein has an apparent size of 140,000, i.e., somewhat less than the major capsid protein. It has been termed the 140K nonglycosylated membrane antigen and is thought to be encoded by nucleotides 1736 to 5689 of the sequence of the B95-8 genome (1) (i.e., BNRF1). The functionally analogous polypeptide from virions of HVS has been designated the 152K (20) or 160K (39–42) polypeptide. By using polyclonal and monoclonal antibodies to the 160K polypeptide of HVS (40, 42) in conjunction with in vitro translation of RNA selected by cloned restriction endonuclease fragments of the genome, the 152/160K product has been mapped to the conventional right-hand end of HVS DNA (*Kpn*I–*Sma*I-E; *Eco*RI–*Sma*I-J–*Eco*RI-K–*Eco*RI-M–*Eco*RI-O) (20; W. Hell, H. Wolf, R. Randall, and R. W. Honess, unpublished

* Corresponding author.

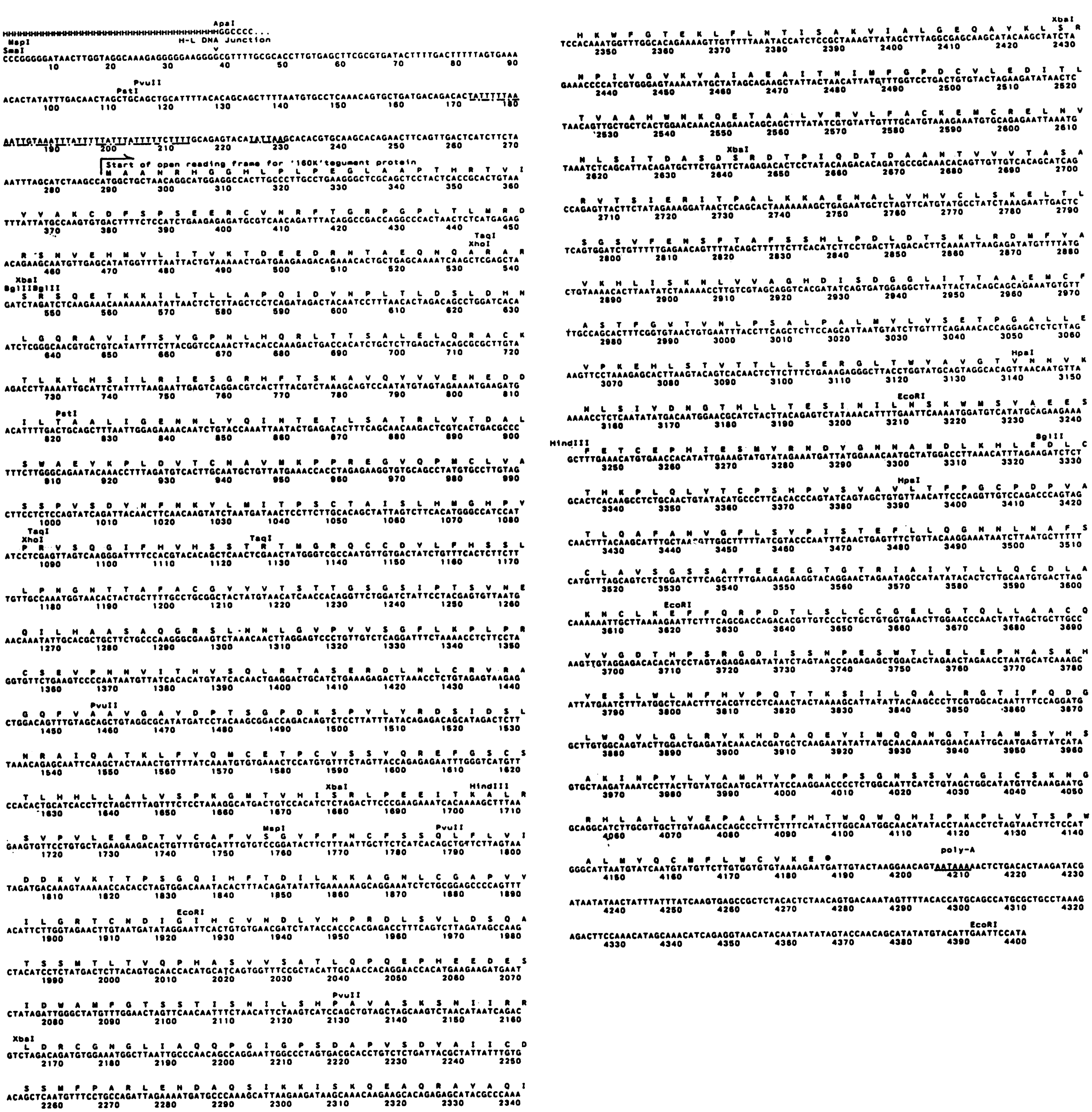

FIG. 1. Sequence of 4,400 bp from the right end of the L-DNA component of HVS strain 11 DNA with the predicted amino acid sequence of the 1,299-residue product of the major ORF. The DNA sequence begins at the *Sma*I site which has conventionally defined the right-hand boundary of the L-DNA component of the HVS genome with the terminal, H-DNA, repeats (the first base of L-DNA is nucleotide 39 of this sequence) and ends 4 nucleotides beyond the *Eco*RI site which separates *Eco*RI-M and *Eco*RI-N (see reference 5 and Fig. 2 for maps of relevant restriction endonuclease cleavage sites). Positions of the recognition sites for selected restriction endonucleases are annotated, as is the position of the first consensus polyadenylation signal (AATAAA, poly-A) on this DNA strand (see Fig. 2).

results). In this paper we present an analysis of the sequence of this region of the HVS genome and provide proof that these functionally analogous proteins are homologous gene products.

MATERIALS AND METHODS

DNA sequencing. The DNA sequence was obtained as part of the determination of the complete sequence of the *Kpn*I–*Sma*I-E (9 kbp) fragment of HVS strain 11 (15) DNA. The fragment was purified from plasmid pWD11 (27) and used to produce three libraries of fragments: (i) random fragments, created by sonication of the *Kpn*I–*Sma*I-E fragment, cloned into the *Sma*I site of M13mp18 (35); (ii) products of partial digestion with *Sau*3A cloned into the *Bam*HI site of M13mp18; and (iii) *Eco*RI fragments (N, M, O, and K) cloned into the *Eco*RI site of M13mp18. All these

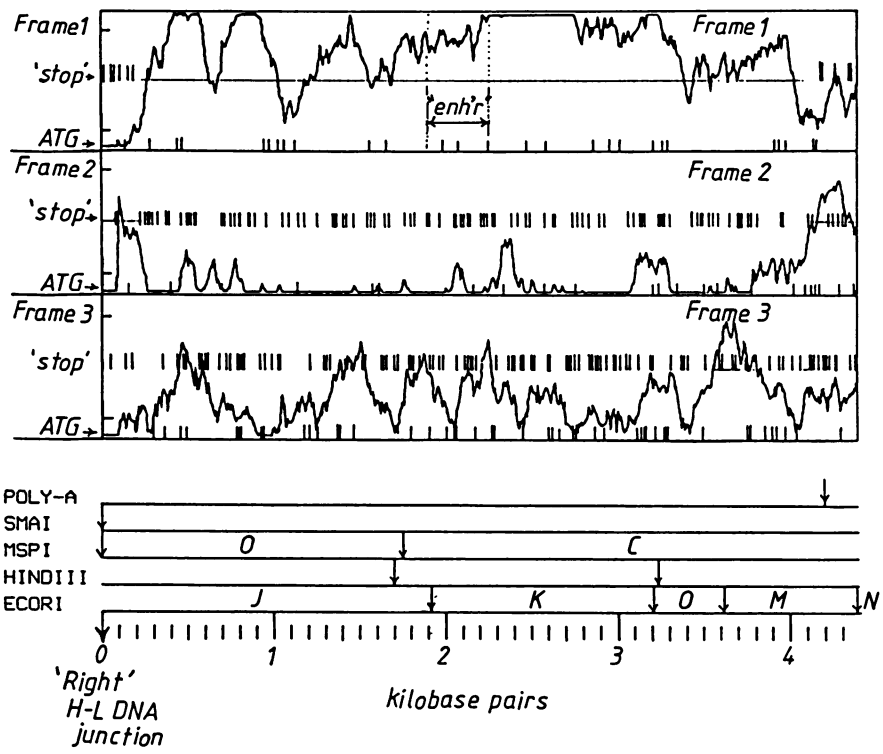

FIG. 2. Measures of the positional base preference (continuous tracings) relative to positions of stop codons and potential initiation codons (ATG) in each of the three reading frames of the sequence given in Fig. 1. The positions of a number of restriction endonuclease cleavage sites and the single polyadenylation (POLY-A) site on this DNA strand are indicated against the scale at the base of the figure. The display was obtained with the ANALYSEQ programs of Staden (49, 50), with the positional base preference averaged over a sliding window of 67 codons. The location of the sequences isolated by Schirm et al. (46) as a substitute for the simian virus 40 enhancer is indicated (enh'r) in frame 1, the reading frame of the 160K protein.

clones were sequenced by the M13 dideoxynucleotide methods of Sanger et al. (44, 45), by using the techniques described in detail by Bankier and Barrell (2), with buffer gradient electrophoresis and adenosine 5'-[^{35}S]thiotriphosphate (>400 Ci/mmol; The Radiochemical Centre, Amersham, U.K.) as the labeled nucleotide (4). The region from nucleotides 1 to 2272 of the sequence given in Fig. 1 was sequenced independently in our two laboratories, including the 377-bp sequence that was previously reported to serve as a substitute for the simian virus 40 enhancer (46) (nucleotides 1896 to 2272 of the present sequence). There was only a single discrepancy between the two sequences (Schirm et al. [46] report a T at nucleotide 2200, whereas a C is given in Fig. 1; the assignment of a C was confirmed by resequencing clones across this region).

TABLE 1. Observed and expected frequencies of dinucleotides

Dinucleotide	No.	% Observed	% Expected[a]
AA	426	9.7	10.0
AC	309	7.0	6.8
AG	326	7.4	5.8
AT	329	7.5	9.1
CA	374	8.5	6.8
CC	164	3.7	4.6
CG	60	1.4	3.9
CT	343	7.8	6.1
GA	255	5.8	5.8
GC	223	5.1	3.9
GG	120	2.7	3.4
GT	210	4.8	5.3
TA	336	7.6	9.1
TC	244	5.5	6.1
TG	302	6.9	5.3
TT	378	8.6	8.2

[a] Given the following mononucleotide frequencies and assuming random associations: A, 1,391 (31.6%); G, 808 (18.4%); C, 941 (21.4%); T, 1,260 (28.6%).

Assembly and analysis of the DNA sequence. Sequences were assembled by using the DB programs of Staden (48), DBUTIL and DBAUTO, and the corrected consensus sequence was analyzed by using the ANALYSEQ programs (49, 50) and the Molecular Genetics and Sequencing (MGS) suite of programs developed by W. Greer and P. Gillett, National Institute for Medical Research, London. Protein compositional analyses, structure prediction (6, 17), and data base searches also involved the use of options in the MGS programs and in ANALYSEP (51). All these computations were performed on a DEC 20-60 computer. Searches with the MGS programs have been performed with issues of public data bases up to issue 8 of the European Molecular Biology Organization nucleic acids data base (April 1986; 6,396 entries) and issue 10 of the NBRF protein sequence data base (August 1986; 4,248 entries) and with local data bases containing translations of all major reading frames of EBV (1) and of published sequences of other herpesvirus genes (9). In addition, searches for homologies to the HVS protein sequence were performed with the FASTP program of Lipman and Pearson (30) and the PROTSEQ library

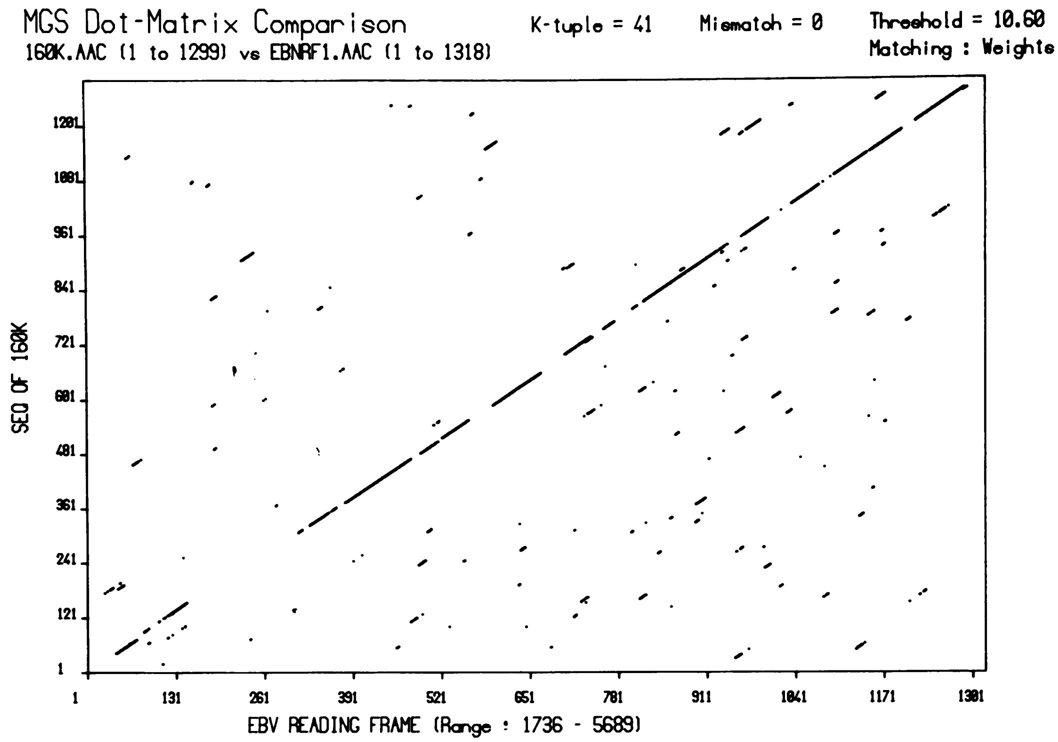

FIG. 3. Comparisons of the predicted amino acid sequences of the 160K polypeptide of HVS (ordinate) and the 140K product of the BNRF1 reading frame of EBV (abscissa; the translation of nucleotide 1736 to 5689 from the EBV sequence of Baer et al. [1]) by a dot-matrix representation of all homologous subsequences which exceed a threshold score. The figure shows the output from a program (MGS dot-matrix) similar to the DIAGON program of Staden (47). The program compares all sequential combinations of *K* residues (*K* = 41 residues in the case illustrated) from one sequence with the other by using a weight matrix (the weight matrix was based on a minimum mutation distance matrix [47], with values ranging from 2 for W versus C to 27 for W versus W) and places a point at each position for which the sum of the weights of paired residues exceeds a value of *K* multiplied by the threshold weight. The threshold weight was set at 10.6 (per residue).

(810,584 residues in 3,557 sequences) on a VAX 11/780 computer at the Laboratory for Molecular Biology, University of Cambridge, Cambridge, U.K. Independent analyses of the properties of the DNA sequence and its predicted protein product in Erlangen were done with the Wisconsin package (12) implemented on a VAX 11/780 computer.

RESULTS

The DNA sequence: composition and general properties. The sequence presented in Fig. 1 comprises 4,400 nucleotides and extends from the *Sma*I site, which has conventionally defined the junction with the H-DNA repeat units at the right of the genome, through the *Eco*RI site, which separates the *Eco*RI M and N fragments of the L-DNA component of HVS strain 11. The sequence has a mean composition of 39.8% G+C, and the occurrence of CpG dinucleotides is much lower than would be expected on the basis of this mean composition and a random distribution between dinucleotides (Table 1: CpG/GpC = 1.4/5.1). An estimate of the significance of the deviation of observed from expected occurrences of CpG was obtained by analyses of randomized samples of the current sequence. From 35 randomized versions of the current sequence, the mean observed occurrences of CpG is the same as that of GpC, i.e., 155, with a standard deviation of 9.4. The number of occurrences of CpG in the real sequence (i.e., 60; Table 1) thus deviates from the expected mean value by more than 10 standard deviations. However, the specific suppression of CpG

TABLE 2. Comparison of predicted amino acid compositions for the 160K polypeptide of HVS and the 140K product of the BNRF1 reading frame of EBV

Amino acid[a]	No. in:		% in:		% 160K/ % BNRF1[b]
	BNRF1	160K	BNRF1	160K	
Asp	55	56	4.17	4.31	1.03
Asn	42	65	3.19	5.00	1.57
Thr	69	96	5.24	7.38	1.41
Ser	91	118	6.90	9.08	1.31
Glu	71	67	5.39	5.15	0.96
Gln	59	51	4.48	3.92	0.82
Pro	91	70	6.90	5.38	0.78
Gly	126	67	9.56	5.15	0.54
Ala	128	104	9.71	8.00	0.82
Cys	25	39	1.90	3.00	1.58
Met	31	26	2.35	2.00	0.85
Val	92	84	6.98	6.46	0.93
Ile	38	68	2.88	5.23	1.81
Leu	153	138	11.61	10.62	0.91
Tyr	32	44	2.43	3.38	1.39
Phe	53	43	4.02	3.31	0.82
Trp	18	13	1.37	1.00	0.73
His	34	45	2.58	3.46	1.34
Lys	23	50	1.75	3.85	2.2
Arg	87	55	6.60	4.23	0.64

[a] Amino acid totals: BNRF1, 1,318; 160K, 1,299. Molecular weights: BNRF1, 142,681; 160K, 143,067.

[b] Major differences in the gross composition of the two polypeptides are underlined.

dinucleotides is a general property of the HVS genome, not a specific feature of this sequence, and will be documented and discussed in detail elsewhere (R. W. Honess, U. A. Gompels, K. R. Cameron, B. G. Barrell, and R. Staden, manuscript in preparation). The positions of stop codons and of potential initiation codons (ATG) in each of the three frames of this DNA strand are summarized in Fig. 2, together with a measure of the positional base preference in each of these three frames. A consistent bias in the positional base preference is a general property of coding sequences (50, 52). The sequence of this DNA strand contains a single major open reading frame (ORF), and a consistent bias in the positional base preference is specific to the region of this ORF (Fig. 2). The sequence of the 1,299 amino acids making up the predicted product of this reading frame is given in Fig. 1. The longest ORF on the opposite strand is 363 bp, and no consistent positional bias is associated with this or the other ORFs on this other strand (not shown). Results comparable to the general measure of positional bias illustrated in Fig. 2 were obtained by measurements of conformity to a table of codon usages derived from other recently characterized genes of HVS (22; data not shown). The expected bias in the use of A+T-rich synonyms of degenerate codons is the main feature of the coding sequence, but superimposed upon this general property is the additional reduction in codons containing the suppressed dinucleotide, CpG (e.g., there are 65 occurrences of CCA or CCT versus 5 occurrences of CCC or CCG and 76 occurrences of TCA or TCT versus 8 occurrences of TCC or TCG).

A region of 377 bp (from nucleotides 1896 to 2272 of the sequence in Fig. 1) from the present sequence was previously isolated from HVS DNA as a functional substitute for the simian virus 40 enhancer in the enhancer-trap assay (46). This sequence is in the middle of the large ORF. If a major part of this sequence served an essential function as an enhancer (or any other *cis*-recognition function) in addition to encoding for protein, we would expect significant departures in composition and codon usage from those observed for portions of the reading frame with no such dual function. No significant departures from the average patterns of dinucleotide frequencies, positional base preference (Fig. 2), or codon usage (not shown) are evident across this region.

A markedly A+T-rich sequence (51 of 58 nucleotides are A or T in the region from nucleotides 172 to 229; Fig. 1) is located some 60 nucleotides 5′ to the major ORF. The first occurrence of the consensus polyadenylation signal, AATAAA, begins 18 nucleotides 3′ to the translational stop codon which ends this frame (Fig. 1 and 2); there are three further occurrences of this signal within the next 600 bp (not shown).

Properties of the protein product. The protein encoded by the major ORF (Fig. 1) is predicted to have a molecular weight of about 143,000 (Table 2), which is in good agreement with experimental estimates of 152,000 to 160,000 for the major protein of the virion tegument (the 160K protein). The protein has an unusually high content of serine and threonine (16.4%).

Searches for homologous protein sequences in a library of the major reading frames predicted from the complete sequence of the Epstein-Barr virus genome (1) revealed a highly significant homology to the product of the BNRF1 reading frame (nucleotides 1736 to 5689) at the left end of EBV. This protein was also the only significant extended homology among sequences of the NBRF/PIR and PROTSEQ databases (QQBE1; probable membrane antigen, p140 of EBV). For example, with the FASTP program of Lipman and Pearson (30), with a *K* tup of 2, the optimized score of the HVS protein versus this EBV protein was 1,200 relative to a mean score (for all other sequences in the PROTSEQ library) of 25.7 and the next highest optimized score of 64. The general extent and quality of the homology of the HVS 160K protein and the EBV gene product can be appreciated from a dot-matrix display of the arrangement of all homologous subsequences which exceed a threshold score (Fig. 3). The two proteins are of very similar size (Table 2; Fig. 3), and homologous subsequences are colinear. The 300 residues from the amino termini of the sequences contain relatively fewer homologous regions than the remainder of the proteins.

```
   1 MAANRHGGHLPLPEGLAAPTHRTVIVVAKCDFSPSEERCVNRFTGRPGPLTLMRDRSNVEHMVLI   65
     |    :   :|:      |     ::         |||     :     | |       | ::|:
   1 MEERGRETQMPVARYGG-PFIMVRLFGQDGEANIQEERLVELLSDPRSALGLDPGPLIAENLLLV   64

  66 TVKTDEEDRNTAEQNQARARSRS-------QETKKILTLLAPQIDVNPLTLDSLDHNLGQRAVIF  123
      ::    |     |  ||            ::  :: |  | :         | :  :      :
  65 ALRGTNNDPRPQRQERARELALVGILLGNGEQGEHLGTESALEA--------SGNNVVVAYGPDW  121

 124 SVGPNLHQRLTTSALELQRACKTLKLHSILRIESGRHFTSKAVQVVVENEDDILTAALIGENNLY  188
        |          | |  |   |::          | :             :|    :   :
 122 MARPSTWSAEIQQFLRLLGATVVLRVEMGRQFGFEVHRSRPSFRQFQAINHLVLFDNALRKYDSG  186

 189 QINTETLSATRLVTDALSWAEYKPLDVTCNAVMKPPREGVQPMCLVASSPVSDVNFNKYLMITPS  253
     |:      |  :              :           | |    :         :   |:: |:
 187 QVAAGFQRALLVAGPETADTRPDLRKLNEWVFGGRAAGGRQLADELKIVSALRDTVSGHLVLQPT  251

 254 CTAISLHMGHPYPRVSQGIFHVHSSTRTMGRQCCDVLF----HSSLLPNGNTTAFACGYVVTSTT  314
      |  : ::     |    : |            |  ||    |  | |  |  |   |:|  : |
 252 ETLDTWKVLSRDTRTAHSLEHGFIHAAGTIQANCPQLFMRRQHPGLFPFVNAIASSLGWVVQTAT  316

 315 GSGSIPTSVNE-QILHAASAQGRSLNNLGVPVVSGFLKPLPRCSEVPNNVITHVS--QLRTASER  376
     | |         |      |        ||||| || : :         :   :   :|
 317 GPGADARAAARRQQAFQTRAAAECHAKSGVPVVAGFVRTINATLKGGEGLQPTMFNGELGAIKHQ  381

 377 DLNLCRVRAGQFVAAVGAVDPTSGPDKSPVLYRDSIDSLNRAIQATKLFVQMCETPCVSSVQREF  441
      |:  |   | ::  :| : | ||    |  |  |         |  ||  :   ||:| | |
 382 ALDTVRVDVGHVLIMLGPFQPWSGLTAPPCPYAESSWAQAAVQTALELFSALVPAPCISGVARPP  446

 442 GSCSTLHHLLALVSPKGMTVHISRLPEEITKALRSVPVLEEDT--VCAFVSGYFFNCFSSQLFLV  504
     |    : ||  ||   |: : :|:||  :   |  :         :  ||| || |   | :|:
 447 GPSAVIEHLGSLVPKGGLLLFLSHLPDDVKDGLGEMGPARATGPGMQQFVSSYFLNPACSNVFIT  511

 505 IDDKVKTTPSGQIHFTDILKKAGNLCGAPVVILGRTCNDIGIHCVNDLYHPRDLSVLDSQATSSM  569
     :  :        :     | :| :: |   |:|| |    |:  ||||  |     :    :
 512 VRQRGEKINGRTVLQA--LGRACDMAGCQHVVLGSTVPLGGLNFVNDLASPVSTAEMMDDFSPFF  574

 570 TLT---VQPHASVVSATLQPQEPHEEDESIDWAMFGTSSTISNILSHPAVASKSNIIRRLDRCGN  631
     |:    :|         |   |      |         | :: ||||| | ||  ::|: ||
 575 TVEFPPIQEEGASSPVPLDVDESMDISPSVELPWLSLESCLTSILSHPTVGSKEHLVRHTDRVSG  639

 632 GLIAQQPGIGPSDAPVSDVAIICDSSMFPARLENDAQSIKKISKQEAQRAVAQIHKWFGTEKLFL  696
     | :|||||:|| | |: ||| :  | ::           :   :        :
 640 GRVAQQPGVGPLDLPLADVAFVAHSQVWTRPGGAPPLPYRTWDRMT------EKLLVSAKPGGEN  698

 697 NTISAKVIALGEQAVKLSRNPIVGVKVAIAEAITNIMFGPDCVLEDITLTVAAHWNKQETAALVR  761
       :|  || |||| ||:| :   | : |:|||: |    |    ||: ||:  | :
 699 VKVSGTVITLGEQGVKVSLDLREGTRLAMAEALLNAACAPILDPEDVLLTLHLHLDPRRADNSAV  763

 762 VLFACKEMCRELNVNLSITDASDSRDTPIQDTDAANTVVVTASARVT-SIERITPALKKAENALV  825
     :            : : :|  | |            ||| : ||    |   ||| | |    |:
 764 MEAMTAASDVARGLGVKLTFGSASCPETGSSASNFMTVVASVSAPGEFSGPLITPVLQKTGSLLI  828

 826 HVCLSKELTLSGSVFENSFTAFSSHLPDLDTSKLRDMFVAVKHLISKNLVVAGHDISDGGLITTA  890
      |         ||:||  |:   :         |:::| ||  |:   :|: |||||||||:|
 829 AVRCGDGKIQGGSLFEQLFSDVATTPRAPEALSLKNLFRAVQQLVKSGIVLSGHDISDGGLVTCL  893

 891 AEMCFASTFGVTVNLPSALPALMYLVSETPGALLEVPKEHLSTVTTLLSERGLTWYAVGTVNNVK  955
      ||  |   |||: :| |   |  :  | || : ||    :  |   |    :    :| |
 894 VEMALAGQRGVTITMPVASDVLPEMFAEHPGLVFEVEERSVGEVLQTLRSMNMYPAVLGRVGEQG  958

 956 NLSIVDNGTH---LLTESINILNSKWMSYAEESFETCEPHIE--SMVRNDYGNNAMDLKHLEDLC 1015
        ::        :| :|: :|   | |:| | :|   |     ||   ||| |
 959 PDQMFEVQHGPETVLRQSLRLLLGTWSSFASEQVECLRPDRINRSMHVSDYGVNEALAVSPLTGK 1023

1016 THKPLQLVTCPSHPVSVAVLTFPGCPDPVATLQAFANVGFLSYPISTEFLLQGNNLNAFSCLAVS 1080
        |  | | |     ||||  ||       | || | | |   :    :     |: :  ||:
1024 NLSPRRLVTEPDPRCQVAVLCAPGTRGHESLLAAFTNAGCLCRRVFFREVRDNTFLDKVVGLAIG 1088

1081 GSSAFEEEGTGTRIAIVTLLQCDLAKNCLKEFFQRPDTLSLCCGELGTQLLAACQVVGDTH-PSR 1144
     |           |  :  :      :: :  |  |||| |:  |||| |:||    || |  |
1089 GVHGARDSALAGRATVALINRFPALRDAILKFLNRPDTFSVALGELGVQVLAGLGAVGSTDNPPA 1153

1145 GDISSNPESWTLELEPNASKHVESLWLNFHVPQTTKSIILQALRGTIFQDGLWQV-LGLRVKHDA 1208
       :  | :   | | ||||  :|| |||  :| || |::|  ||| :    :    ||| :
1154 PGVEVNVQRSPLILAPNASGMFESRWLNISIPATTSSVMLRGLRGCVLPCWVQGSCLGLQFTNLG 1218

1209 QEVIMQQNGTIAMSVHSAKINPVLVAMHVPRNPSGNSSVAGICSKNGRHLALLVEPALSFHTWQW 1273
       |::|    ||  :||   : : :|| |||||:    :||:||::|||||||  | |    |||
1219 MPVVLQNAHQIACHFHSNGTDAWRFAMNVPRNPTEQGNIAGLCSRDGRHLALLCDPSLCTDFWQW 1283

1274 QHIP----KPLVTSPWALMYQCMFLWCVK    1298-HVS
     :|||    :|   ||| ||:|   ||  ::
1284 EHIPPAFGHPTGCSPWTLMFQAAHLWSLR    1312-EBV
```

FIG. 4. Alignments of the amino acid sequences of the 160K protein of HVS (upper lines) and the 140K product of the BNRF1 reading frame of the EBV genome (lower lines). The alignment shown was obtained by using the algorithm of Wilbur and Lipman (30) as implemented in the MGS suite of programs. Identical residues in the two aligned sequences are interconnected by dashes, and similar residues are connected by dots; gaps introduced to give the alignments shown are indicated by dashes.

The amino acid compositions of the two proteins are compared in Table 2, and an alignment of the two protein sequences is given in Fig. 4. The HVS protein contains significantly more lysine and less arginine, more isoleucine and less leucine and valine, and more asparagine and cystine

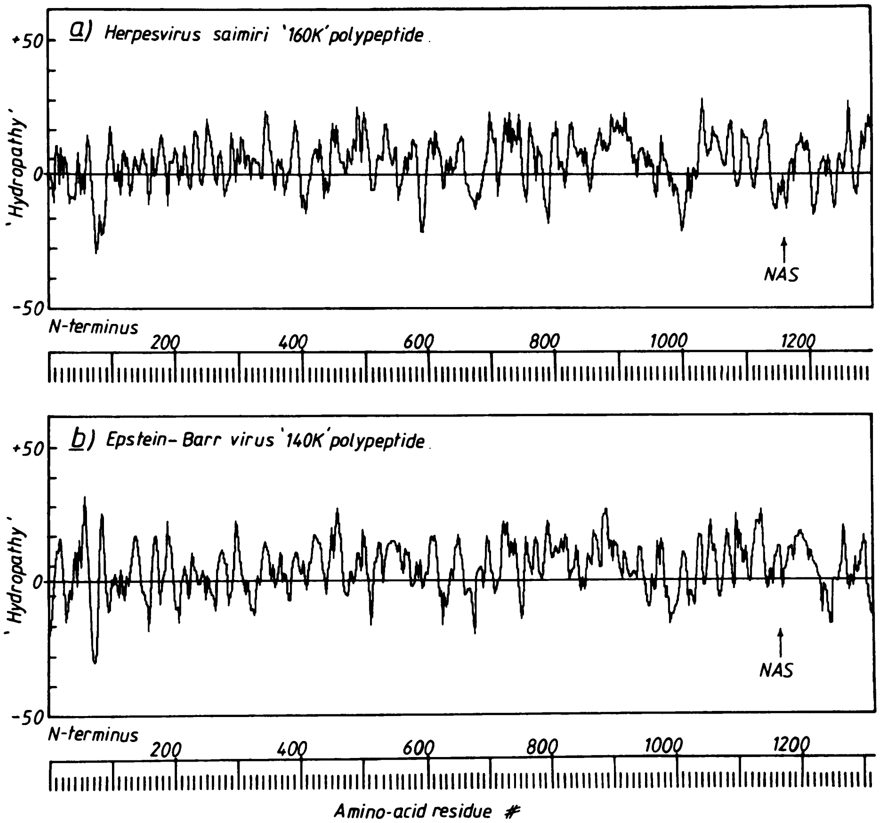

FIG. 5. Comparison of changes in hydrophobicities (ordinates, hydropathy) of (a) the predicted sequence of the 160K protein of HVS and (b) the predicted sequences of the 140K protein of EBV. The graphs are generated by the ANALYSEP program of Staden (51), with the hydropathy values of Kyte and Doolittle (29) with a window size of 11 residues. The ordinate is scaled from +50 (hydrophobic/hydropathic) to −50 (hydrophilic). The position of a conserved asparagine-alanine-serine (NAS) sequence (28) is indicated.

and less glycine and proline than its EBV homolog. In general, these differences are predicted as a result of a relative reduction in amino acids encoded by G+C-rich codons and an increase in those encoded by A+T-rich codons in the HVS sequence. The alignment of the two sequences (Fig. 4) allows a more objective display of the arrangement and extent of homologous subsequences, and it is evident that the most homologous regions are in the C-terminal halves of the two proteins. The overall homology of the alignment is 37%. Amino acid residues 537 to 662 of the HVS sequence are specified by the DNA sequences isolated in the simian virus 40 enhancer trap (46). Residues 537 to 552 and residues 605 to 659 of the HVS protein are clearly conserved with respect to the sequence of the EBV counterpart. If a part of this region of the HVS genome is serving an additional (noncoding) function in the virus, its most likely location would be between these conserved regions (i.e., nucleotides 1944 to 2099 of the sequence in Fig. 1).

Available evidence on the roles of the 160K protein of HVS and the EBV gene product suggests that these gene products are not integral membrane proteins but may associate with the underside of the lipid envelope of the virion. Both proteins are either very poorly glycosylated or nonglycosylated. A comparison of the changes in local hydrophobicity for the two proteins is illustrated in Fig. 5. Both proteins are relatively hydrophobic, and the patterns of local variation in hydrophobicity correlate well, particularly over the C-terminal two-thirds of the molecules. However, neither protein has a signal sequence or a sufficiently extended region of hydrophobic residues to indicate a cotranslational or posttranslational membrane insertion sequence (32). Analyses of the predicted secondary structures of both proteins (6, 17) indicate an overall preponderance of β-sheet over α-helix and overall conservation of secondary structure, as expected from the extent of sequence conservation (data not shown). Without further experimental data on the secondary and tertiary structures of these proteins, such predictions do not currently provide the basis for realistic modeling of these structures.

DISCUSSION

In this paper we have shown that the sequences at the right end of the HVS genome contain a single major ORF and that the product of this reading frame is homologous to the product of the BNRF1 reading frame at the left end of EBV. The 160K proteins of HVS previously mapped to this region of the HVS genome (20) is synthesized at maximal rates late in a productive infection, its synthesis is sensitive to inhibitors of virus DNA synthesis (36), and it is incorporated as a major component of the virion of HVS (25, 39–42). Recent studies of its assembly and topography in the mature virion has confirmed its role as a tegument component, not exposed through the lipid bilayer, and have shown that the neutralizing properties of a rabbit polyclonal serum to the 160K protein were due to minor contaminating antibodies to virion glycoproteins (40, 42; M. R. Gopal and R. W. Honess, unpublished results). The EBV homolog is also expressed as a late gene (24) and is a major structural protein of virions of EBV (13, 14). The DNA sequences upstream of the EBV gene (bases 1 to 1735 [1]) have a mean nucleotide composition of 47% G+C relative to 61.4% G+C for the BNRF1 reading frame, and we have noted the unusually A+T-rich sequences which are upstream of the HVS gene. Both genes are immediately preceded by potential TATA box homologies and followed by polyadenylation signals (AATAAA). Thus, the regulatory class and the relative positions of presumptive *cis*-regulatory sequences, as well as the nature of the encoded proteins, have been conserved in these two gamma herpesviruses within DNA sequences which differ in composition by some 20% G+C. More significantly, in both these viruses, these genes are in similar positions and the same orientation with respect to the signals which determine genome termini, and there is no obvious homolog of the gene among reading frames predicted from the complete sequence of the alpha herpesvirus varicella-zoster virus (9).

Comparisons of the genetic organization of the alpha herpesviruses and, in particular, herpes simplex virus and varicella-zoster virus, have established that there is a general colinearity of homologous genes in this subfamily (8–11). Genes which are clearly homologous to a number of the structural and nonstructural genes identified in these alpha herpesviruses have also been recognized in the genome of the gamma herpesvirus EBV (1, 7–10, 18, 19, 31, 33, 34, 37, 38). Although there is conservation in the relative order and arrangement of homologous reading frames within portions of these genomes, the global gene order observed in the alpha herpesviruses is not conserved in the genome of EBV. In addition, there are many regions of the genomes of varicella-zoster virus and EBV which appear to specify proteins with no obvious homology to proteins of the other virus. One of the minimal events required to relate the order of genes of the alpha herpesviruses to genes which are recognizably homologous to EBV is a nonconservative transposition of the signals which determine the genome termini in these two virus subgroups.

The results of this paper clearly show that at least one of the termini of HVS and of EBV are at similar positions with respect to a gene which is common to these gamma herpesviruses but is not recognized in the alpha herpesviruses. This observation suggests that despite differences in their mean nucleotide compositions, gamma herpesvirus genomes share features of their genetic repertoire and organization which are not observed in representatives of other herpesvirus subgroups.

ACKNOWLEDGMENTS

We thank Wolfgang Hell for providing copies of his unpublished results and W. Greer, P. Gillett, and J. Green, National Institute for Medical Research, and R. Staden, Laboratory for Molecular Biology, University of Cambridge, Cambridge, United Kingdom, for providing their sequence analysis (MGS, ANALYSEQ, and ANALYSEP) and assembly (DBUTIL and DBAUTO) programs. We also thank B. G. Barrell for access to the FASTP data base search programs implemented on a VAX.

Work in Erlangen was supported by the Deutsche Forschungsgemeinschaft.

LITERATURE CITED

1. **Baer, R., A. T. Bankier, M. D. Biggin, P. L. Deininger, P. J. Farrell, T. J. Gibson, G. Hatful, G. S. Hudson, S. C. Satchwell, C. Seguin, P. S. Tuffnell, and B. G. Barrell.** 1984. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature (London) **310**:207–211.
2. **Bankier, A. T., and B. G. Barrell.** 1983. Shotgun DNA sequencing, p. 1–33. *In* R. A. Flavell (ed.), Techniques in the life sciences, vol. B5. Elsevier/North Holland Scientific Publishers, Ltd., Limerick, Republic of Ireland.
3. **Bankier, A. T., W. Dietrich, R. Baer, B. G. Barrell, F. Colbere-Garapin, B. Fleckenstein, and W. Bodemer.** 1985. Terminal repetitive sequences in herpesvirus saimiri virion DNA. J. Virol. **55**:133–139.
4. **Biggin, M. D., T. Gibson, and G. F. Hong.** 1983. Buffer-gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80**:3963–3965.
5. **Bodemer, W., E. Knust, A. Angermuller, and B. Fleckenstein.** 1984. Immediate-early transcription of herpesvirus saimiri. J. Virol. **51**:452–457.
6. **Chou, P. Y., and G. Fasman.** 1978. Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. Relat. Areas Mol. Biol. **47**:45–148.
7. **Costa, R. H., K. G. Draper, T. J. Kelly, and E. K. Wagner.** 1985. An unusual spliced herpes simplex virus type 1 transcript with sequence homology to Epstein-Barr virus DNA. J. Virol. **54**:317–328.
8. **Davison, A. J., and D. J. McGeoch.** 1986. Evolutionary comparisons of the S segments in the genomes of herpes simplex virus type 1 and varicella-zoster virus. J. Gen. Virol. **67**:597–611.
9. **Davison, A. J., and J. E. Scott.** 1986. The complete DNA sequence of varicella-zoster virus. J. Gen. Virol. **67**:1759–1816.
10. **Davison, A. J., and J. E. Scott.** 1986. DNA sequence of the major capsid protein gene of herpes simplex virus type 1. J. Gen. Virol. **67**:2279–2286.
11. **Davison, A. J., and N. M. Wilkie.** 1983. Location and orientation of homologous sequences in the genomes of five herpesviruses. J. Gen. Virol. **64**:1927–1942.
12. **Devereux, J., P. Haeberli, and O. Smithies.** 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:387–395.
13. **Dolyniuk, M., R. Pritchett, and E. Kieff.** 1976. Proteins of Epstein-Barr virus. I. Analysis of the polypeptides of purified enveloped Epstein-Barr virus. J. Virol. **17**:935–949.
14. **Dolyniuk, M., E. Wolff, and E. Kieff.** 1976. Proteins of Epstein-Barr virus. II. Electrophoretic analysis of the polypeptides of the nucleocapsid and the glucosamine- and polysaccharide-containing components of enveloped virus. J. Virol. **18**:289–297.
15. **Falk, L. A., L. G. Wolfe, and F. Deinhardt.** 1972. Isolation of herpesvirus saimiri from blood of squirrel monkeys (*Saimiri sciureus*). J. Natl. Cancer Inst. **48**:1499–1505.
16. **Fleckenstein, B., and R. C. Desrosiers.** 1982. Herpesvirus saimiri and herpesvirus ateles, p. 253–332. *In* B. Roizman (ed.), The herpesviruses, vol. 1. Plenum Publishing Corp., New York.
17. **Garnier, J., D. J. Osguthorpe, and B. Robson.** 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. **120**:97–120.
18. **Gibson, T., P. Stockwell, M. Ginsburg, and B. Barrell.** 1984. Homology between two EBV early genes and the HSV ribonucleotide reductase and 38K genes. Nucleic Acids Res. **12**:5087–5099.
19. **Gompels, U., and A. Minson.** 1986. The properties and sequence of glycoprotein H of herpes simplex virus type 1. Virology **153**:230–247.
20. **Hell, W., S. Modrow, and H. Wolf.** 1985. Mapping of herpesvirus saimiri proteins on the viral genome: proteins dependent and not dependent on viral DNA synthesis. J. Virol. **56**:414–418.
21. **Honess, R. W.** 1984. Herpes simplex and 'the herpes complex': diverse observations and a unifying hypothesis. J. Gen. Virol. **65**:2077–2107.
22. **Honess, R. W., W. Bodemer, K. R. Cameron, H.-H. Niller, B. Fleckenstein, and R. E. Randall.** 1986. The A+T rich genome of herpesvirus saimiri contains a highly conserved gene for thymidylate synthase. Proc. Natl. Acad. Sci. USA **83**:3604–3608.
23. **Honess, R. W., and D. H. Watson.** 1977. Unity and diversity in the herpesviruses. J. Gen. Virol. **37**:15–37.
24. **Hudson, G. S., A. T. Bankier, S. C. Satchwell, and B. G. Barrell.**

1985. The short unique region of the B95-8 Epstein-Barr virus genome. Virology **147**:81–98.

25. **Keil, G., B. Fleckenstein, and W. Bodemer.** 1983. Structural proteins of herpesvirus saimiri. J. Virol. **47**:463–470.
26. **Kieff, E., T. Dambaugh, W. King, M. Heller, A. Cheung, V. van Santen, M. Hummel, C. Beisel, and S. Fennewald.** 1982. Biochemistry of Epstein-Barr virus, p. 105–150. *In* B. Roizman (ed.), The herpesviruses, vol. 1. Plenum Publishing Corp., New York.
27. **Knust, E., S. Schirm, W. Dietrich, W. Bodemer, E. Kolb, and B. Fleckenstein.** 1983. Cloning of herpesvirus saimiri DNA fragments representing the entire L-region of the genome. Gene **25**:281–289.
28. **Kornfeld, R., and S. Kornfeld.** 1985. Assembly of asparagine-linked oligosaccharides. Annu. Rev. Biochem. **54**:631–644.
29. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.
30. **Lipman, D. J., and W. R. Pearson.** 1985. Rapid and sensitive protein similarity searches. Science **227**:1435–1441.
31. **Littler, E., J. Zeuthen, A. A. McBride, E. Trost Sorensen, K. L. Powell, J. E. Walsh-Arrand, and J. R. Arrand.** 1986. Identification of an Epstein-Barr virus-coded thymidine kinase. EMBO J. **5**:1959–1966.
32. **McGeoch, D. J.** 1985. On the predictive recognition of signal peptide sequences. Virus Res. **3**:271–286.
33. **McGeoch, D. J., and A. J. Davison.** 1986. DNA sequence of the herpes simplex virus type 1 gene encoding glycoprotein gH, and identification of homologues in the genomes of varicella-zoster virus and Epstein-Barr virus. Nucleic Acids Res. **14**:4281–4292.
34. **McGeoch, D. J., A. Dolan, and M. C. Frame.** 1986. DNA sequence of the region in the genome of herpes simplex virus type 1 containing the exonuclease gene and neighbouring genes. Nucleic Acids Res. **14**:3435–3448.
35. **Messing, J.** 1983. New M13 vectors for cloning. Methods Enzymol. **101**:20–78.
36. **O'Hare, P., and R. W. Honess.** 1983. Identification of a subset of herpesvirus saimiri polypeptides synthesized in the absence of virus DNA replication. J. Virol. **46**:279–283.
37. **Pellett, P. E., M. D. Biggin, B. Barrell, and B. Roizman.** 1985. Epstein-Barr virus genome may encode a protein showing significant amino-acid and predicted secondary structure homology with glycoprotein B of herpes simplex virus. J. Virol. **56**:807–813.
38. **Quinn, J. P., and D. J. McGeoch.** 1985. DNA sequence of the region in the genome of herpes simplex virus type 1 containing the genes for DNA polymerase and the major DNA binding protein. Nucleic Acids Res. **13**:8143–8163.
39. **Randall, R. E., and R. W. Honess.** 1980. Proteins of herpesvirus saimiri: identification of two virus polypeptides released into the culture medium of productively infected cells. J. Gen. Virol. **51**:445–449.
40. **Randall, R. E., and R. W. Honess.** 1982. Proteins specified by herpesvirus saimiri: purification and properties of a single polypeptide which elicits virus-neutralizing antibody. J. Gen. Virol. **58**:149–161.
41. **Randall, R. E., R. W. Honess, and P. O'Hare.** 1983. Proteins specified by herpesvirus saimiri: identification and properties of virus-specific polypeptides in productively infected cells. J. Gen. Virol. **64**:19–35.
42. **Randall, R. E., C. Newman, and R. W. Honess.** 1984. Isolation and characterization of monoclonal antibodies to structural and nonstructural herpesvirus saimiri proteins. J. Virol. **52**:872–883.
43. **Roizman, B.** 1982. The family Herpesviridae: general description, taxonomy, and classification, p. 1–23. *In* B. Roizman (ed.), The herpesviruses, vol. 1. Plenum Publishing Corp., New York.
44. **Sanger, F., A. R. Coulson, B. G. Barrell, A. J. H. Smith, and B. A. Roe.** 1980. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. J. Mol. Biol. **143**:161–178.
45. **Sanger F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
46. **Schirm, S., F. Weber, W. Schaffner, and B. Fleckenstein.** 1985. A transcription enhancer in the herpesvirus saimiri genome. EMBO J. **4**:2669–2674.
47. **Staden, R.** 1982. An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. Nucleic Acids Res. **10**:2951–2961.
48. **Staden, R.** 1982. Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Res. **10**:4731–4751.
49. **Staden, R.** 1984. Graphic methods to determine the function of nucleic acid sequences. Nucleic Acids Res. **12**:521–538.
50. **Staden, R.** 1984. Measurements of the effects that coding for protein has on a DNA sequence and their use for finding genes. Nucleic Acids Res. **12**:551–567.
51. **Staden, R.** 1986. The current status and portability of our sequence handling software. Nucleic Acids Res. **14**:217–231.
52. **Staden, R., and A. D. McLachlan.** 1982. Codon preference and its use in identifying protein coding regions in long DNA sequences. Nucleic Acids Res. **10**:141–156.