Volume 14 Number 10 1986 Nucleic Acids Research

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

Duncan J.McGeoch* and Andrew J.Davison[1]

MRC Virology Unit, Institute of Virology, University of Glasgow, Church Street, Glasgow G11 5JR, UK

Received 13 March 1986; Accepted 22 April 1986

ABSTRACT

We have determined the sequence of herpes simplex virus type 1 DNA around the previously mapped location of sequences encoding an epitope of glycoprotein gH, and have deduced the structure of the gH gene and the amino acid sequence of gH. The unprocessed polypeptide is predicted to contain 838 amino acids, and to possess an N-terminal signal sequence and a C-terminal transmembrane sequence. Temperature-sensitive mutant tsQ26 maps within the predicted gH coding sequence. Homologous genes were identified in the genomes of two other herpesviruses, namely varicella-zoster virus and Epstein-Barr virus.

INTRODUCTION

The virion of herpes simplex virus (HSV) possesses an outer envelope consisting of a lipid bilayer in which are embedded a number of glycoprotein species. By the early 1980s it appeared, following a period of some confusion, that both serotypes of HSV encoded four membrane glycoprotein species, termed gB, gC, gD and gE (for review, see ref. 1). However, recent work, in particular the application of monoclonal antibody techniques and of DNA sequence analysis, has detected other glycoprotein species or has shown that further glycoproteins may be encoded in the HSV genome. Thus, an HSV-2 glycoprotein named gG or g92K has been described (2,3,4), encoded by a gene in the short unique region (U_S: see Figure 1) of the HSV-2 genome. In HSV-1, sequence analysis of the U_S region has indicated the presence of three genes thought to encode "extra" glycoproteins (5,6), one of which is the HSV-1 equivalent of HSV-2 gG (7; also, unpublished data). Lastly, Buckmaster et al. (8) used a monoclonal antibody to define a new glycoprotein of HSV-1, termed gH, whose gene mapped to a position in the long unique

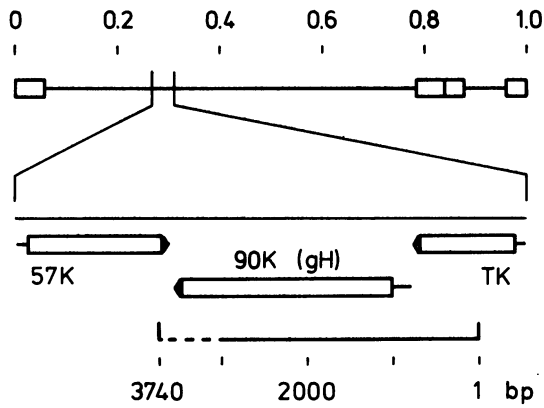


Figure 1. Organization of the gH gene region in the genome of HSV-1. The upper part of the figure depicts the prototype HSV-1 genome, with numbering in fractional map units. The long and short unique sequences are shown as solid lines, with major repeat elements as open boxes. The middle part of the figure expands a 6 kb region from 0.268 to 0.312 map units, to show the layout of 57K, 90K (gH) and TK genes. Positions and orientations of transcripts are indicated, with predicted coding regions as open boxes. The lower part of the figure shows the mapping bracket for the gH epitope (0.282 to 0.308 map units; 8), with numbering as in the sequence listing of Figure 2. The position of the mapping bracket's left end is uncertain in terms of the HSV-1 DNA sequence (see text), and the region of uncertainty is indicated as a dashed line.

region (U_L: see Figure 1) distinct from other glycoprotein genes.

This paper is concerned with the identity and structure of the gH gene, which lies in a part of the HSV-1 genome for which we have determined the sequence. We have located the gH gene and have deduced the encoded amino acid sequence of gH. In addition, we have identified corresponding genes encoded by the alphaherpesvirus varicella-zoster virus (VZV) and by the gammaherpesvirus Epstein-Barr virus (EBV).

MATERIALS AND METHODS

The DNA sequences of plasmid cloned copies of HSV-1 strain 17 restriction fragments were determined by the M13/chain terminator system as described (5,9,10). For the sequence reported in this paper, fragments used were EcoRI g cloned in

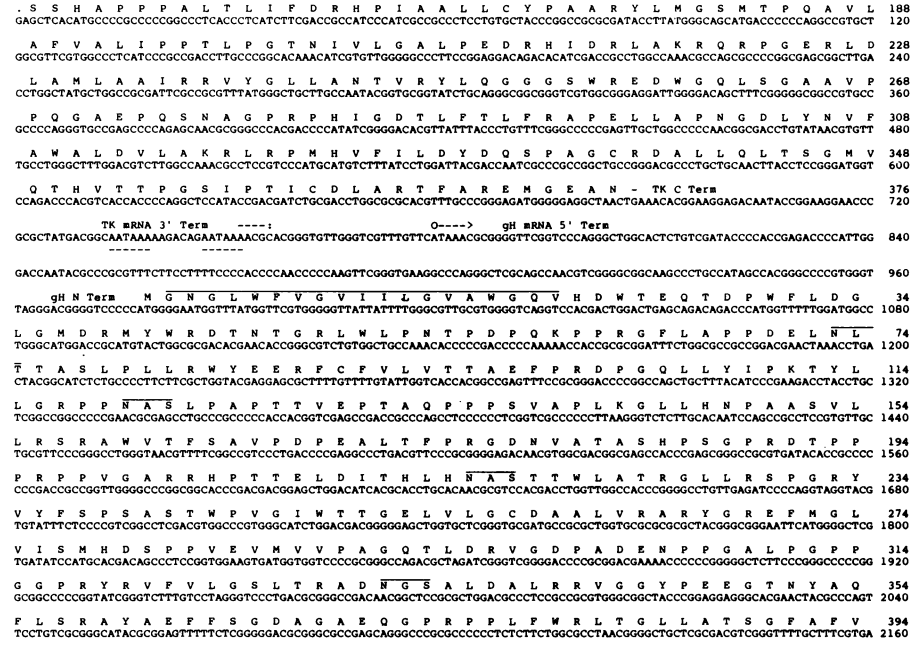
pACYC184 (from V.G. Preston) and BamHI p cloned in pAT153 (11). Computer handling and interpretation of sequence data used a PDP11/44 under RSX11M as described (5,10).

RESULTS

(a) Identification of the HSV-1 gene encoding gH

The monoclonal antibody (LP11) used to define HSV-1 gH is type specific (that is, it is not active against HSV-2) (8). This enabled Buckmaster et al. (8) to locate the portion of the HSV-1 genome encoding the epitope recognized by LP11 to a 4 kb region in U_L, between 0.282 and 0.308 map units, by assaying activity of LP11 against a reference set of intertypic recombinants. As part of a large scale DNA sequence analysis of the HSV-1 genome, we have determined the sequence of this region, and this is shown in Figure 2, as 3740 bp of composition 65.7% G+C. For a reason which will become apparent, the sequence is listed as the strand oriented 5' to 3', right to left, on the genome diagram of Figure 1. The right boundary of the gH epitope mapping bracket is marked by an SstI site in HSV-1 (11) (residue 1 in Figure 2). The left boundary is defined by a KpnI site in HSV-2 (12), and so cannot at present be placed with exactitude on the HSV-1 sequence, but lies between HSV-1 KpnI sites at residues 2969 and 3735 in Figure 2. Together with published mRNA mapping data and sequence studies (13,14,15), the sequence enabled the deduction of gene arrangement and of amino acid sequences of encoded proteins. As shown in outline in Figure 1, and explicitly in Figure 2, the region contains all or part of three genes. At its right extremity lies the downstream portion of the leftward transcribed thymidine kinase (TK) gene (13,14). To the left of the target region there is a rightward transcribed gene (15) encoding a protein of predicted M_r 57,638 (here called 57K), of unknown function (our unpublished data). The mapping bracket may just include a small part of the 57K coding sequence. Finally, between the TK and 57K genes there is a leftward transcribed gene (16), encoding a protein of predicted M_r 90,360 (now called 90K), and this gene is completely, or almost completely, within the mapping bracket.

The well-studied TK gene can at once be dismissed as a candidate for encoding gH. The 57K gene can also be reasonably excluded, since the encoded protein does not possess any visible signal sequence or transmembrane segment. In addition, only the sequence encoding the 19 C-terminal amino acids can lie within



the mapping bracket, even on the most favourable interpretation. In contrast, the 90K protein encoded by the centrally placed gene possesses a hydrophobic N-terminal region which could comprise a signal sequence (6), and a second hydrophobic region, adjacent to the C-terminus, which could be a transmembrane anchor sequence (see below). The protein sequence also contains 7 potential N-glycosylation sites. The unprocessed polypeptide has a predicted M_r appropriate for a precursor of glycosylated gH, which has an estimated M_r of 110,000 to 120,000 (8,17). Finally, it is already known that the 90K gene is transcribed, late in lytic infection (16).

It is clear from these arguments that the 90K gene is an excellent candidate for encoding gH and also that no real alternative is discernible within the mapped locus. We therefore conclude that we have located the gH gene. While this stops short of a formal identification, nonetheless, from the viewpoint of DNA sequence interpretation the conclusion appears well justified. In the following sections we describe the sequence of the gH gene and characteristics of the protein, and identify corresponding genes in the genomes of two other herpesviruses.

(b) The gH gene and polypeptide

The mRNA species which we now believe to encode gH was previously characterized by Sharp et al. (16) as an abundant, late transcript of approximately 3 kb, whose 5'-terminus was located 23 residues 3' to the 3'-terminus of TK mRNA (that is, at residue 779 in Figure 2). As those authors pointed out, this implies that the promoter for the 3 kb transcript must overlap the 3'-terminus of the TK gene. We now consider that the mRNA's

Figure 2. DNA sequence of the gH gene region in the genome of HSV-1. The sequence is shown for the leftward 5'-3' strand only for the gH gene region, as indicated in Figure 1. This sequence was obtained using plasmid-cloned fragments BamHI p (residues 1 to 2305) and EcoRI g (residues 1785 to 3740). The 5'-terminus of gH mRNA is indicated as "O---->", and the 3'-terminus of TK mRNA (13,14,15), together with predicted 3'-termini for gH and 57K mRNAs, as "----:". Polyadenylation associated sequences AATAAA are underlined. Predicted amino acid sequences are shown for gH and for the C-terminal portions of TK and the 57K protein. In the gH amino acid sequence, hydrophobic regions representing probable signal and transmembrane sequences are overlined, as are possible N-glycosylation sites.

3'-terminus should be adjacent to one or both of the appropriately positioned polyadenylation associated sequences AATAAA at residues 3490 and 3531. This would give the "3 kb" RNA a length, excluding poly(A), of 2740 to 2780 residues. Downstream of these termination sites there is an intergenic region of 90 to 130 residues, followed by the 3'-terminus of the 57K gene (see Figure 2).

Within the "3 kb" mRNA region, the first potential translation initiation codon is at residue 978, and this opens a reading frame of 838 codons, terminating with TAA at 3492, which is thought to encode gH. The predicted amino acid sequence exhibits an uncharged region of 20 residues at its N-terminus. We have previously shown by criteria of length and hydrophobicity that this region probably comprises a signal sequence for translation on membrane bound ribosomes (6). Adjacent to the C-terminus there is another stretch of uncharged amino acids (residues 975 to 824), followed by several basic residues, which we presume to comprise a transmembrane anchor region (18). These two hydrophobic sequences are illustrated by the "hydropathy" plot of Figure 3. Within the proposed external domain (up to residue 794) there are seven occurrences of potential N-glycosylation sites (N-S and N-T; ref. 20). The 838 codon open reading frame would encode a polypeptide of M_r 90,360. From proposals for prediction of the site of cleavage by signal peptidase (21,22), it is most likely that cleavage would occur after residue 18, leaving an 820 residue polypeptide of M_r 88,489.

In 1983, mapping of the temperature-sensitive mutation in HSV-1 strain KOS tsQ26 to a small region adjacent to the TK gene was reported (23). On the sequence listing of Figure 2, the tsQ26 mapping bracket is bounded by the PvuII site at 1290 and the EcoRI site at 1785, a region wholly within gH coding sequence. Thus, it is now clear that tsQ26 is a mutation in the gH gene, demonstrating that gH is an essential protein, at least in tissue culture infection.

We have compared the predicted gH amino acid sequence with sequences of other HSV-1 glycoproteins, including gB (24), gC (25), gD (5), gE (5) and three other species, encoded in the

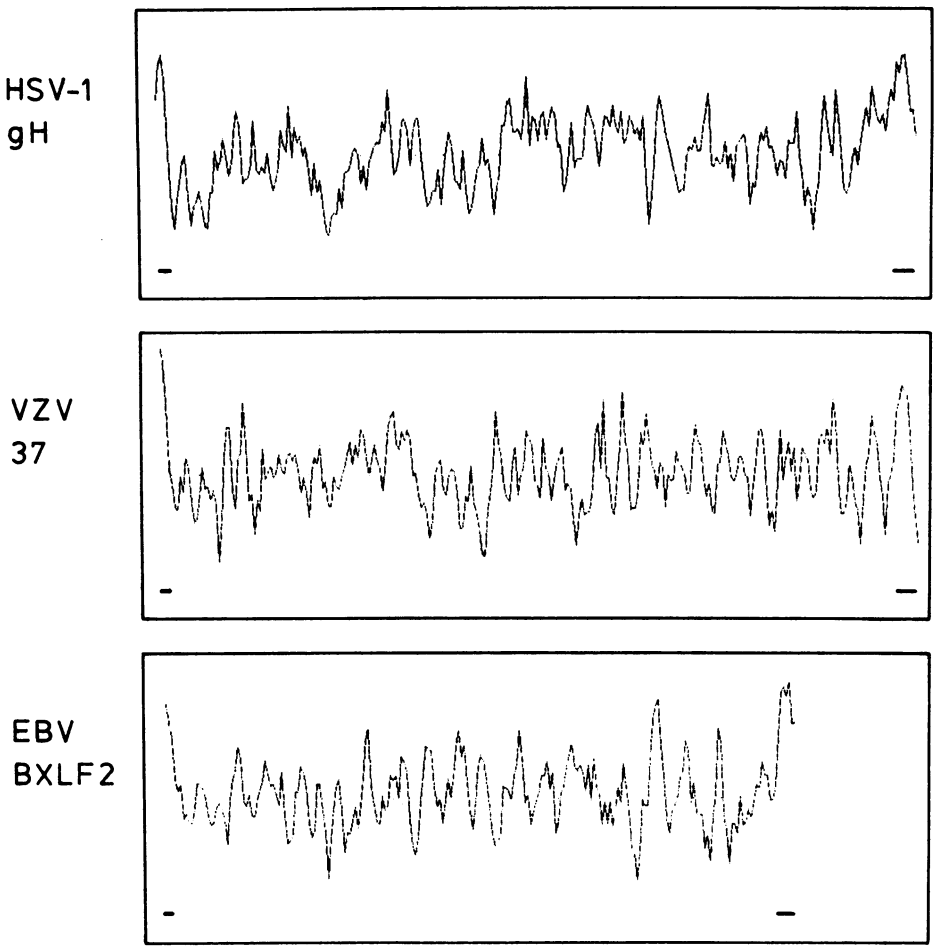


Figure 3. Hydropathy plots for predicted amino acid sequences of herpesvirus glycoproteins. Scans of the local hydrophobic and hydrophilic characters are shown for the predicted amino acid sequences of gH and the corresponding VZV and EBV proteins (see text). For each, the x-axis represents a set of 11-residue windows on the sequence, with successive windows incremented by 3 residues, and the N-terminus at the left. The y-axis represents the hydropathy sum (19) for each window (scale: -40 to +40). High values represent hydrophobic regions, and low values hydrophilic regions. Proposed signal sequence and transmembrane sequence hydrophobic regions are underlined.

short unique region (5,6,7). However, no significant similarities were found.

(c) Identification of VZV and EBV counterparts of the gH gene

HSV and VZV both belong to the Alphaherpesvirinae

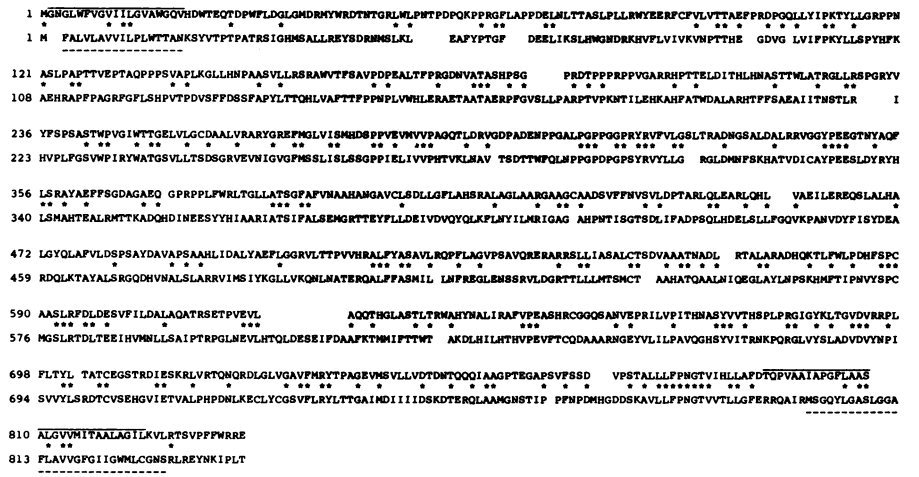


Figure 4. Comparison of HSV-1 gH and VZV gene 37 proteins. The predicted amino acid sequences for gH and VZV gene 37 protein (28) were aligned by the program of Taylor (29), using default parameters. The gH sequence is the upper, and pairs of identical residues are indicated by asterisks. Gaps introduced by the program are shown as blanks. Proposed signal sequences and transmembrane sequences are overlined or underlined.

sub-family, although they differ substantially in their DNA sequences and in details of genome organization (26,27). The complete sequence of the VZV genome has recently been determined (28), and this has allowed us to identify a counterpart to the HSV-1 gH gene, namely VZV gene 37. These genes occupy corresponding positions in each genome, and comparison of the two predicted amino acid sequences shows clear homology. In the alignment shown in Figure 4, the sequences exhibit 25% matching. The VZV gene 37 polypeptide would contain 841 amino acids in its unprocessed form. Like HSV-1 gH, there are hydrophobic regions at the N-terminus and near the C-terminus, thought to be the signal sequence and transmembrane sequence, respectively (Figures 3 and 4). There are ten potential N-glycosylation sites.

In 1984, Baer et al. (30) published the complete genome sequence of EBV, which is classified as a member of the Gammaherpesvirinae sub-family (26). This is only distantly

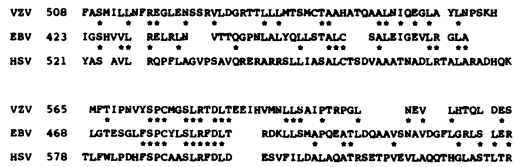


Figure 5. Comparison of EBV BXLF2 amino sequence with the two alphaherpesvirus sequences. Alignments are shown for portions of the BXLF2, VZV gene 37 protein and gH sequences. Identical pairs of residues between BXLF2 and either of the others are marked by asterisks. Gaps introduced to obtain alignment are shown as blanks. This figure was constructed from computed alignments (29) of BXLF2 with each of the other sequences separately. It does not represent an overall optimal alignment of all three sequences.

related to the Alphaherpesvirinae, but several EBV genes have now been shown to have counterparts in HSV (31,32,33,34). From the complete VZV genome sequence (28), a number of homologous VZV and EBV genes have been identified, and an overall relationship between the gene arrangements of the two genomes has emerged (A.J. Davison and P. Taylor, in preparation). This showed that VZV gene 37 had a probable counterpart in the EBV reading frame BXLF2 (30), in terms of genome positions. BXLF2 would encode a protein of 706 amino acids. We have evaluated relations between the BXLF2 amino acid sequence, and the sequences of HSV-1 gH and VZV gene 37 protein. The N-terminal half, approximately, of the BXLF2 sequence showed little or no homology with the others by the procedures used, but regions in the C-terminal portions were recognizably related, although generally weakly. Alignment required introduction of many small gaps. The most convincing homology is at residues 475 to 487 in the BXLF2 sequence, and this is shown in Figure 5, together with flanking sequences. This is also one of the regions most conserved between the sequences of gH and VZV gene 37 protein

Figure 6. Terminal regions of EBV BXLF2 protein.
The N-terminal 50 residues and C-terminal 50 residues are listed for the BXLF2 protein (30), with overlining to indicate proposed signal sequence and transmembrane sequence.

(Figure 4). As shown in Figures 3 and 6, BXLF2 also possesses candidate signal and transmembrane sequences. There are five potential N-glycosylation sites. In summary, it is clear that this EBV gene is related to the two alphaherpesvirus genes, although widely diverged, and encodes a membrane-inserted protein, presumably a virion glycoprotein. A promoter for the gene has been identified, which is active late in the replicative cycle (30).

DISCUSSION

We conclude from these studies that we have identified the HSV-1 gene for gH, and that the protein has a standard arrangement of N-terminal signal sequence, a number of possible N-glycosylation sites and a C-terminal membrane anchor region. The function of HSV-1 gH is presently unknown, but its importance was indicated by the findings that monoclonal antibody LP11 could neutralize virus infectivity and also, uniquely, inhibit plaque formation when added after the start of infection (8). Our conclusion that the previously mapped tsQ26 mutation (23) lies in the gH gene shows that gH is essential, and the finding that the gH gene has counterparts in VZV and in EBV, which is not the case for several of the other HSV glycoproteins, could also argue a basic functional role for gH and its homologues.

Biochemical and immunological studies have distinguished four VZV glycoproteins, and the genes for three of these have been identified (35,36,37). These, however, do not include gene 37. The complete DNA sequence contains five probable glycoprotein genes (28). It seems likely that gene 37 encodes the glycoprotein designated gpIII by Davison et al. (37), but we have no direct evidence on this.

In the case of EBV, present knowledge regarding virion glycoproteins is quite limited. Three species, termed gp350, gp220 and gp85, have been recognized (38,39). From the complete genome sequence, Baer et al. (30) proposed the existence of five glycoprotein genes. For only two of these, encoding the related gp350 and gp220 species, have the corresponding glycoproteins been identified (40,41). Another gene encodes a protein

homologous to HSV-1 gB (34). BXLF2 was not suggested as a possible glycoprotein gene by Baer et al., so it now appears that EBV may encode six glycoproteins.

ACKNOWLEDGEMENTS

We acknowledge the assistance of A. Dolan, D. McNab and J. Scott.

*To whom correspondence should be addressed

[1]Present address: Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA

REFERENCES

1. Spear,P.G. (1985) In The Herpesviruses, Roizman,B. Ed. Vol 3, pp. 315-356, Plenum Press, New York.
2. Marsden,H.S., Buckmaster,A., Palfreyman,J.W. Hope,R.G. and Minson,A.C. (1984) J. Virol. 50, 547-554.
3. Roizman,B., Norrild,D., Chan,C. and Pereira,L. (1984) Virology 133, 242-247.
4. Oloffson,S., Lundstrom,M., Marsden,H.S., Jeansson,S. and Vahlne, A. (1986) J. Gen. Virol., in press.
5. McGeoch,D.J., Dolan,A., Donald,S. and Rixon,F.J. (1985) J. Mol. Biol. 181, 1-13.
6. McGeoch,D.J. (1985) Virus Res. 3, 271-286.
7. Frame,M.C., Marsden,H.S. and McGeoch,D.J. (1986) J. Gen. Virol., in press.
8. Buckmaster,E.A., Gompels,U. and Minson,A. (1984) Virology 139, 408-413.
9. Bankier,A.T. and Barrell,B.G. (1983) In Techniques in the Life Sciences, Flavell,R.A. Ed., Vol B508, pp.1-34, Elsevier, Ireland.
10. McGeoch,D.J., Dolan,A., Donald,S. and Brauer,D.H.K. (1986) Nucleic Acids Res. 14, 1727-1745.
11. Sanders,P.G., Wilkie,N.M. and Davison,A.J. (1982) J. Gen. Virol. 63, 277-295.
12. Marsden,H.S., Stow,N.D., Preston,V.G., Timbury,M.C. and Wilkie,N.M. (1978) J. Virol. 28, 624-642.
13. McKnight,S.L. (1980) Nucleic Acids Res. 8, 5949-5964.
14. Wagner,M.J., Sharp,J.A. and Summers,W.C. (1981) Proc. Nat. Acad. Sci. U.S.A. 78, 1441-1445.
15. Wagner,E.K. (1985) In The Herpesviruses, Roizman,B. Ed., Vol 3, pp.45-104, Plenum Press, New York.
16. Sharp,J.A., Wagner,M.J. and Summers,W.C. (1983) J. Virol. 45, 10-17.
17. Showalter,S.D., Zweig,M. and Hampar,B. (1981) Infect. Immun. 34, 684-692.
18. Wickner,W.T. and Lodish,H.L. (1985) Science 230, 400-407.
19. Kyte,J. and Doolittle,R.F. (1982) J. Mol. Biol. 157, 105-132.
20. Hubbard,S.C. and Ivatt,R.J. (1981) Annu. Rev. Biochem. 50, 555-583.
21. Von Heijne,G. (1983) Eur. J. Biochem. 133, 17-21.

22. Perlman,D. and Halvorson,H.O. (1983) J. Mol. Biol. 167, 391-409.
23. Weller, S.K., Aschman,D.P., Sacks,W.R. Coen,D.M. and Schaffer,P.A. (1983) Virology 130, 290-305.
24. Bzik,D.J., Fox,B.A., DeLuca,N.A. and Person,S. (1984) Virology 133, 301-314.
25. Draper,K.G., Costa,R.H., Lee,G.T.-Y., Spear,P.G. and Wagner,E.K. (1984) J. Virol. 51, 578-585.
26. Matthews,R.E.F. (1982) Intervirology 17, 1-199.
27. Davison,A.J. and McGeoch,D.J. (1986) J. Gen. Virol., in press.
28. Davison,A.J. and Scott,J.E. (1986) J. Gen. Virol., in press.
29. Taylor,P. (1984) Nucl. Acids Res. 12, 447-456.
30. Baer,R., Bankier,A.T., Biggin,M.D., Deininger,P.L., Farrell,P.J., Gibson,T.J., Hatfull,G., Hudson,G.S., Satchwell,S.C., Seguin,C., Tuffnell,P.S. and Barrell,B.G. (1984) Nature 310, 207-211.
31. Gibson, T., Stockwell,P., Ginsburg,M. and Barrell,B. (1984) Nucl. Acids Res. 12, 5087-5099.
32. Costa,R.H., Draper,K.G., Kelly,T.J. and Wagner,E.K. (1985) J. Virol. 54, 317-328.
33. Quinn,J.P. and McGeoch,D.J. (1985) Nucl. Acids Res. 13, 8143-8163.
34. Pellett,P.E., Biggin,M.D., Barrell,B. and Roizman,B. (1985) J. Virol. 56, 807-813.
35. Davison,A.J., Waters,D.J. and Edson,C.M. (1985) J. Gen. Virol. 66, 2237-2242.
36. Ellis,R.W., Keller,P.M., Lowe,R.S. and Zivin,R.A. (1985) J. Virol. 53, 81-88.
37. Davison,A.J., Edson,C.M., Ellis,R.W., Forghani,B., Gilden,D., Grose,C., Keller,P.M., Vafai,A., Wroblewska,Z. and Yamanishi,K. (1986) J. Virol., in press.
38. Strnad,B.C., Schuster,T., Klein,R., Hopkins,R.F., Witmer,T., Neubauer,R.H. and Rabin,H. (1982) J. Virol. 41, 258-264.
39. Edson,C.M. and Thorley-Lawson,D.A. (1983) J. Virol. 46, 547-556.
40. Biggin,M., Farrell,P.J. and Barrell,B.G. (1984) EMBO J. 3, 1083-1090.
41. Beisel,C., Tanner,J., Matsuo,T., Thorley-Lawson,D., Kezdy,F. and Kieff,E. (1985) J. Virol. 54, 665-674.