The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome

Andrew J. Davison,¹ Aidan Dolan,¹ Parvis Akter,¹ Clare Addison,¹ Derrick J. Dargan,¹ Donald J. Alcendor,² Duncan J. McGeoch¹ and Gary S. Hayward²

Correspondence

Andrew Davison
a.davison@vir.gla.ac.uk

¹MRC Virology Unit, Institute of Virology, Church Street, Glasgow G11 5JR, UK

²Molecular Virology Laboratories, Oncology Center, Johns Hopkins School of Medicine, Baltimore, MD 21231, USA

Received 24 May 2002

Accepted 18 September 2002

The gene complement of wild-type human cytomegalovirus (HCMV) is incompletely understood, on account of the size and complexity of the viral genome and because laboratory strains have undergone deletions and rearrangements during adaptation to growth in culture. We have determined the sequence (241 087 bp) of chimpanzee cytomegalovirus (CCMV) and have compared it with published HCMV sequences from the laboratory strains AD169 and Toledo, with the aim of clarifying the gene content of wild-type HCMV. The HCMV and CCMV genomes are moderately diverged and essentially collinear. On the basis of conservation of potential protein-coding regions and other sequence features, we have discounted 51 previously proposed HCMV ORFs, modified the interpretations for 24 (including assignments of multiple exons) and proposed ten novel genes. Several errors were detected in the published HCMV sequences. We presently recognize 165 genes in CCMV and 145 in AD169; this compares with an estimate of 189 unique genes for AD169 made in 1990. Our best estimate for the complement of wild-type HCMV is 164 to 167 genes.

INTRODUCTION

Human cytomegalovirus (HCMV; human herpesvirus 5) is ubiquitous and largely inapparent, but poses a risk of serious disease to those lacking a competent immune system, such as neonates, transplant patients and sufferers from AIDS (reviewed in Pass, 2001). HCMV is the prototype of subfamily *Betaherpesvirinae*, and is the most complex of the eight human herpesvirus species. HCMV is isolated routinely on human fibroblast cell lines, and several strains in common laboratory use, such as AD169 and Towne, were derived by multiple passages on such cells (reviewed in Mocarski & Tan Courcelle, 2001).

The linear, double-stranded DNA genome of AD169 comprises two covalently linked segments (L and S), each consisting of a unique region (U_L and U_S) flanked by an inverted repeat (TR_L and IR_L , TR_S and IR_S), yielding the overall genome configuration $TR_L-U_L-IR_L-IR_S-U_S-TR_S$ (reviewed in Mocarski & Tan Courcelle, 2001). In addition,

the genome is terminally redundant, possessing a short region (the *a* sequence) as a direct repeat at the termini and also in inverse orientation at the IR_L-IR_S junction. Some genomes contain tandemly reiterated copies of the *a* sequence at these locations. U_L and U_S can invert relative to each other by recombination between inverted repeats in replicating DNA, resulting in four equimolar genome arrangements in virion DNA. The complete DNA sequence of AD169 was published in a seminal paper by Chee *et al.* (1990), and at that time was the largest viral genome sequence available. The total genome size was 229 354 bp, with U_L being 166 972 bp, U_S 35 418 bp, R_L (a collective term for TR_L and IR_L) 11 247 bp, R_S (TR_S and IR_S) 2524 bp and the *a* sequence (part of R_L and R_S in the sizes given above) 578 bp.

As a primary criterion for identifying protein-coding regions, Chee *et al.* (1990) focused on open reading frames (ORFs) of 100 or more contiguous amino acid-encoding codons that overlapped larger ORFs by no more than 60 % of their length. Smaller ORFs were identified in remaining gaps on the basis of appropriately located transcriptional elements, amino acid sequence similarity to other AD169 ORFs or genes from other organisms, structural or functional motifs in amino acid sequences, and codon bias.

The GenBank accession number of the CCMV sequence reported in this paper is AF480884, and that for our third-party annotation of the HCMV AD169 sequence is BK000394. Details of the updated interpretations of the HCMV Toledo and Towne sequences are available from the author for correspondence.

This process led to the identification of a total of 208 potentially protein-coding ORFs, a number which, when duplications in R_L and known splicing were taken into account, reduced to 189 unique genes. Given the limitations of the criteria, Chee *et al.* (1990) noted that some ORFs in this set might not actually encode proteins, and that small or highly spliced genes might have been missed.

Chee *et al.* (1990) rightly viewed their picture of HCMV gene content as the best achievable at the time, and anticipated that it would be modified. Indeed, subsequent experimental mapping and sequence reinterpretation have changed it in several ways. Three sequence differences (presumably errors) have been recognized, the first extending the 5'-end of ORF UL102 (Smith & Pari, 1995), the second also in UL102 (Smith & Pari, 1995), and the third extending the 3'-end of US28 (Neote *et al.*, 1993). Two groups (Dargan *et al.*, 1997; Mocarski *et al.*, 1997) found that certain stocks of AD169 possess an additional 929 bp at a location in U_L , resulting in extensions of UL42 and UL43. Dargan *et al.* (1997) also provided a reinterpretation of UL41, concluding that an alternative reading frame is likely to be the protein-coding ORF in this region. Several ORFs listed by Chee *et al.* (1990) have been modified by evidence for splicing, including UL111A, UL118 and UL119 (Rawlinson & Barrell, 1993), and UL33 (Davis-Poynter *et al.*, 1997). One original ORF, UL22, has been replaced by a spliced gene involving an overlapping reading frame (Rawlinson & Barrell, 1993). Complex splicing patterns have been described for some ORFs, such as US3 (Rawlinson & Barrell, 1993).

Parallel to these refinements to interpretation of the AD169 sequence has come realization that this highly passaged strain lacks several genes present in other isolates. Cha *et al.* (1996) discovered an extra 15 kbp at the right end of U_L in a low passage strain (Toledo), adding 19 ORFs to the HCMV gene complement. They also noted that Towne has a less extensive deletion in the same region. These observations indicate that genetic loss is due to selection imposed by passage in human fibroblasts, and taking into account an expansion of R_L concomitant with the deletion in AD169 (Prichard *et al.*, 2001), would place the genome size of Toledo in the region of 235 kbp. Toledo has become a widely used low passage isolate, but it is apparent from comparisons with other clinical isolates that even this strain may not be representative of wild-type HCMV (Cha *et al.*, 1996; Prichard *et al.*, 2001). These findings imply that no laboratory strain can be taken as genetically complete and since HCMV has not yet been sequenced directly from clinical material and the interpretation of the coding potential of AD169 is still being refined, a full picture of the gene content of wild-type HCMV is not yet available.

A powerful approach to improving the interpretation of a sequence is to compare it with a relative, on the basis that most genuine protein-coding regions will have been conserved during evolution, whereas spurious features

such as non-functional ORFs will not. This has substantially aided definitions of the gene contents of human herpesviruses 6 and 7 (HHV-6 and HHV-7; Megaw *et al.*, 1998), equid herpesviruses 1 and 4 (EHV-1 and EHV-4; Telford *et al.*, 1998) and herpes simplex virus types 1 and 2 (HSV-1 and HSV-2; Dolan *et al.*, 1998). Since the pioneering work of Chee *et al.* (1990), several other members of the *Betaherpesvirinae* have been sequenced, including murine cytomegalovirus (MCMV; Rawlinson *et al.*, 1996), rat cytomegalovirus (RCMV; Vink *et al.*, 2000), tupaiid herpesvirus 1 (TuHV-1; Bahr & Darai, 2001), HHV-6 (Gompels *et al.*, 1995; Dominguez *et al.*, 1999; Isegawa *et al.*, 1999) and HHV-7 (Nicholas, 1996; Megaw *et al.*, 1998), but all are too distant to be of much use in evaluating the coding potential of HCMV. In this paper, we report the genome sequence of chimpanzee cytomegalovirus (CCMV), the closest known relative of HCMV, and reassess the gene layout in HCMV.

METHODS

Cells and virus. A throat swab was obtained by Richard Heberling at the Southwest Foundation for Research and Education, San Antonio, Texas, USA, from a male chimpanzee (*Pan troglodytes*; number 4x148) born in New York on 31 December 1975. CCMV was isolated initially on chimpanzee lung fibroblasts and grown thereafter to passage 16 on human fibroblast cell lines (HFF and MRC5) before storage in 1983. The virus was recently subjected to four additional passages on HFF cells in the laboratory of G. S. H. and titrated. It was not plaque purified.

Preparation of CCMV DNA. Five 175 cm² flasks each containing 10⁷ HFF cells at 80 % confluency were infected with CCMV at an m.o.i. of 0.1 and incubated for 20 days until CPE was complete. Infected cell medium was clarified by centrifugation at 10 000 g for 20 min at 4 °C. Cell-released virus was pelleted from the supernatant by centrifugation through a 5 ml cushion of 15 % (w/v) sucrose in PBS at 70 000 g for 1 h at 4 °C. Virus was resuspended in 400 µl of 50 mM Tris/HCl pH 7.5, 100 mM NaCl, 8 mM MgSO₄, 0.1 % (w/v) gelatin, and treated with 10 µg DNase I µl⁻¹ for 20 min at 37 °C. Virions were lysed by adding SDS to 1 % (w/v), EDTA (pH 8) to 25 mM and proteinase K to 1 µg ml⁻¹, and incubating overnight at 50 °C. After performing phenol/chloroform extraction, the DNA was ethanol precipitated and resuspended in TE (10 mM Tris/HCl pH 7.5, 1 mM EDTA). The yield was 20 µg. The quality of the preparation was confirmed by restriction endonuclease digestion followed by agarose gel electrophoresis and ethidium bromide staining.

DNA sequencing. An M13 library of CCMV sequences was prepared. Virion DNA (5 µg) was sheared randomly by sonication, and fragments (500–1000 bp) were end-repaired and cloned into M13mp19 (Davison & Telford, 1994). Recombinant plaques were picked into 10 µl TE in the wells of 96-well round-bottomed microtitre plates using sterile cocktail sticks. An overnight culture of *Escherichia coli* XL-1 Blue (Stratagene) grown in 2YT broth (85 mM NaCl; 1 %, w/v, bactopeptone; 1 %, w/v, yeast extract) was diluted 1:100 (v/v) in 2YT broth, and 250 µl was added to each well. The plates were covered with rigid lids and incubated with shaking at 37 °C for 6 h in a humidified benchtop incubator. Bacteria were pelleted at 1500 g for 15 min, and 200 µl of the supernatants was transferred to the wells of fresh plates containing 25 µl of 20 % (w/v) PEG, 2.5 M NaCl. The plates were covered with adhesive lids, inverted to mix and incubated overnight at 4 °C. Precipitated

M13 was pelleted at 1500 g for 15 min and the supernatants discarded by inverting the plates. The inverted plates were drained by placing on tissues and centrifuging briefly at 150 g. Bacteriophage was disrupted by adding 40 µl of 4 M NaI in TE, and the plates were covered with rigid lids and shaken in a benchtop incubator at room temperature for 15 min. Ethanol (100 µl) was added to each well, and the contents of the wells were transferred to 96-well PCR plates. The plates were covered with rubber lids, inverted to mix, and incubated at room temperature for 15 min. DNA was pelleted at 2600 g for 30 min and the supernatants were discarded by inverting the plates. The inverted plates were drained by placing on tissues and centrifuging briefly. DNA pellets were washed by adding 100 µl of 95 % (v/v) ethanol to each well, centrifuging the plates briefly, discarding the supernatants and draining the plates as described above. M13 templates were sequenced in an ABI PRISM 377 instrument according to the manufacturer's instructions, using 96-lane gels.

Sequence analysis. The sequence database was compiled from electropherograms using Pregap4 and Gap4 (Staden *et al.*, 2000) and Phred (Ewing & Green, 1998; Ewing *et al.*, 1998). Gaps were closed and regions of difficulty resolved by PCR using specific primers and sequencing the products. The final, edited sequence was subjected to thorough manual checking against the electropherograms. The sequence was analysed using the GCG suite of programs (Genetics Computer Group, Madison, WI, USA) and the Ptrans sequence translation program (Taylor, 1986). Comparisons were carried out with published sequences for the complete AD169 genome (Chee *et al.*, 1990; accession no. X17403, 229 354 bp) and the regions at the right ends of Toledo and Towne U_L (Cha *et al.*, 1996; accession nos U33331 and U33332, 18 535 and 4844 bp).

Identification of genome termini. The genome termini of CCMV were located approximately from similarities to those of AD169, and then mapped experimentally. Briefly, CCMV DNA was treated with T4 DNA polymerase in the presence of the four dNTPs to produce flush ends, and ligated to a partially double-stranded adaptor blocked at the exterior 3'-end (the cDNA adaptor in the Clontech Marathon kit). Each terminus was identified by PCR using a primer (AP1 in the Marathon kit) annealing to the single-stranded region of the adaptor plus a CCMV-specific primer annealing approximately 150 bp from the relevant terminus, cloning the products into pGEM-T (Promega) and sequencing.

Investigation of potential errors in HCMV sequences. Regions of a few hundred base pairs encompassing potential errors in the AD169 or Toledo sequences identified by comparisons with the CCMV sequence were amplified by PCR from infected cell DNA and cloned into pGEM-T. The inserts in at least three independent plasmids were sequenced for each locus. In certain experiments, an AD169 cosmid and DNA extracted from HCMV in the urine of a congenitally infected child (i.e. not passaged in cell culture) were amplified.

RESULTS AND DISCUSSION

Characteristics of the CCMV sequence

The CCMV sequence as represented in the database was determined to an average redundancy of 8.9, and 97 % was obtained on both strands. It contained incomplete copies of TR_L and TR_S at the termini, since random sequences from these elements merged with the corresponding internal repeats. The genome sequence was reconstructed utilizing knowledge of the terminal sequences derived by PCR. Although it represents a consensus from virus that was not plaque purified, few instances of microheterogeneity

were detected. The only previously available CCMV sequence data originated from the major immediate-early promoter (Chan *et al.*, 1996).

PCR experiments using primers mapping on either side of the *a* sequence demonstrated the presence of genomes with a single *a* sequence at the IR_L-IR_S junction, but did not convincingly detect multiple copies (data not shown). However, the existence of reiterated *a* sequences in some genomes was supported by the random sequence data. The sequence across the junction of two *a* sequences contained an additional base pair in comparison with the sequence formed by hypothetical joining of flush-ended genome termini. This was most simply interpreted as indicating an unpaired nucleotide at each genome 3'-terminus. The ultimate residue in the CCMV sequence represents the unpaired nucleotide at the 3'-end of the upper strand.

The size of the CCMV genome with a single *a* sequence at each terminus and at the IR_L-IR_S junction is 241 087 bp. This is consistent with the size measured by electron microscopy (Swinkels *et al.*, 1984). The genome components are: U_L, 199 351 bp; U_S, 35 753 bp; R_L, 687 bp; R_S, 2453 bp; *a* sequence (part of R_L and R_S), 297 bp. The CCMV genome has a G+C content of 61.7 %, 4 percentage points greater than that of AD169. PCR experiments using a primer in IR_S near the internal *a* sequence plus a primer at the left or right end of U_L indicated that U_L is present in either orientation in virion DNA (data not shown). Attempts to investigate inversion of U_S by a similar approach failed, presumably because R_S is larger than R_L. However, restriction endonuclease mapping experiments on appropriate clones from a complete bacteriophage λ library of the genome confirmed the presence of each orientation of U_L and U_S in approximately equal abundance (data not shown). The CCMV genome is thus similar in structure to that of AD169, except that R_L is considerably smaller, as is thought to be the case in wild-type HCMV (Prichard *et al.*, 2001).

Comparison of the CCMV and HCMV sequences

McGeoch *et al.* (2000) carried out a phylogenetic analysis of the proteins encoded by several well conserved CCMV genes sequenced in this study. The degree of relationship accords with the notion that the two viruses evolved with their hosts, with a divergence date approximating that of their host lineages (i.e. 5–6 million years ago). CCMV is the closest known relative of HCMV.

Fig. 1 shows a two-dimensional comparison of the AD169 and CCMV sequences. It is evident that the genomes are related and overall closely collinear, with only a few exceptions of substantial size to this generalization. The largest distinct region of differing organization (around 180–190 kbp in AD169) represents the net outcome of the CCMV sequence possessing about 19 kbp that is missing from the right end of the AD169 U_L (but present in Toledo; Cha *et al.*, 1996) and the presence in AD169

Fig. 1. Matrix plot demonstrating collinearity between the CCMV and AD169 genome sequences. The diagram was computed using GCG Compare (stringency of 24 matches in a moving window of 30 nucleotides) and plotted using GCG Dotplot.

of a larger R_L element than in CCMV (or in Toledo; Prichard *et al.*, 2001) due to internal duplication of the left end of the genome to form part of IR_L . The other prominent difference (around 94–99 kbp in AD169) corresponds to the origin of lytic DNA replication. The two sequences are strongly diverged here, although features of their base compositions across this locus remain broadly similar, with A + T-rich stretches flanking a G + C-rich core. Two-dimensional comparisons of the genomes at higher stringency (not shown) emphasized that large-scale sequence similarity is highest in the central part of U_L and declines toward the genome termini. When subsections of the sequences were compared at higher resolution, many local additions and deletions were apparent that are not visible on the scale of Fig. 1. With these complications, it is not feasible to express overall identity of the aligned genome sequences as a precise single figure. As an indication, UL54 (encoding DNA polymerase in the conserved central region of U_L) gave around 80 % identity; this compares with 90 % identity between the DNA polymerase genes of HSV-1 and HSV-2.

Investigation of potential errors in HCMV sequences

In addition to the three corrections to the published AD169 sequence reported since 1990, our comparisons led to the identification of frameshifts in the AD169 and Toledo sequences. PCR amplification from AD169- or Toledo-infected cell DNA followed by sequencing confirmed that three of these loci differed from the versions published (data not shown). Two differences mapped in the AD169 UL15 ORF, which is not conserved in CCMV.

These involve insertion of single G residues after nucleotides 21955 and 22003, resulting in disruption of UL15 and gain on the other strand of another ORF (UL15A) that is conserved in CCMV. An additional G residue was also identified after nucleotide 211535 in viral DNA and an appropriate cosmid, causing 5'-extension of US22. An additional G residue after nucleotide 8885 in Toledo resulted in 5'-extension of UL145. We also confirmed the two nucleotide substitutions identified in AD169 UL102 by Smith & Pari (1995).

These differences probably reflect errors in the original sequences. They were located initially by comparing candidate coding regions, and it is possible that other errors are present in the AD169 and Toledo sequences in non-coding regions or diverged coding regions.

Gene content of CCMV and HCMV

Developing a picture of the gene content of a large sequence is an involved process, as no set of criteria is perfect. Our primary basis for defining the gene content of CCMV, and reassessing that of HCMV, was the expectation that protein-coding regions should be conserved between viruses exhibiting a moderate degree of evolutionary divergence. We built on previous analyses, but discounted ORFs in one genome that lack positional and sequence counterparts in the other, except where they represent insertions in relation to flanking genes or bioinformatic or functional data indicated otherwise.

Fig. 2 shows the CCMV gene arrangement, with protein-coding regions coloured according to whether they are conserved in the *Alpha*-, *Beta*- and *Gammaherpesvirinae* (core genes) or not (non-core genes). Subsets of non-core genes are related to each other in gene families as indicated. Fig. 3 shows the AD169 gene arrangement, with protein-coding regions now coloured according to changes from the gene layout described by Chee *et al.* (1990). The process of gene identification was largely straightforward but, as expected, left a residuum of uncertainty. We consider that the AD169 gene content presented in Fig. 3 is a substantial improvement over previous versions, but anticipate that further adjustments will be necessary.

We have retained the HCMV gene nomenclature system of Chee *et al.* (1990), and applied it to homologous genes in CCMV. Extension of this system to HCMV genes discovered subsequently is problematic, since different authors have used different approaches. In instances where an altered protein-coding region still incorporates part of the original ORF, the original name has been retained (e.g. UL128). In instances where an original ORF has been replaced by another protein-coding region in a similar location, an additional letter has been added (e.g. UL15A). Previously unidentified protein-coding regions have been named similarly, in logical order (e.g. UL148A). We have used a decimal system for genes whose coding regions correspond to portions of larger, overlapping ORFs. The

CCMV

Fig. 2. Layout of genes in the CCMV genome. The scale is in kbp. R_L and R_S are shown in a thicker format than U_L and U_S . Protein-coding regions are indicated by coloured arrows grouped according to the key, with gene nomenclature below. Introns are shown as narrow white bars. Genes corresponding to those in AD169 R_L and R_S are given their full nomenclature, but the UL and US prefixes have been omitted from UL2–UL157 (10–200 kbp) and US1–US34A (204–237 kbp). Colours differentiate between genes on the basis of conservation in the *Alpha*-, *Beta*- and *Gammaherpesvirinae* (core genes), with subsets of non-core genes grouped into gene families.

Toledo

only example of this in Figs 2 and 3 is UL80.5, which encodes the major capsid scaffolding protein. Other proteins that may be expressed through the use of additional promoters, alternative splicing or lack of splicing, such as those from immediate early genes UL123, UL122, US3 and UL37 (Stenberg *et al.*, 1984, 1985, 1989; Rawlinson & Barrell, 1993; Tenney *et al.*, 1993; Goldmacher *et al.*, 1999), have been omitted for the time being, pending confirmation of functionality during HCMV infection. We predict that UL28 is spliced in HCMV and CCMV to presently unidentified, upstream protein-coding exons, as appears also to be the case for the HHV-6 and HHV-7 homologues (Megaw *et al.*, 1998).

In our approach to determining the gene contents of CCMV and HCMV, virus-specific genes constitute special cases. CCMV lacks counterparts of HCMV UL1, UL111A and UL3. UL1 is a member of the RL11 glycoprotein family. UL111A encodes an interleukin-10 homologue in HCMV and other primate cytomegaloviruses (Kotenko *et al.*, 2000; Lockridge *et al.*, 2000). UL3 is the most marginal gene to be retained in AD169. CCMV contains four genes not present in the AD169 and Toledo sequences: UL146A, UL155, UL156 and UL157. UL146A is related to an adjacent gene, UL146, which in HCMV encodes an α -chemokine (Penfold *et al.*, 1999). UL157 is also related to UL146, and UL155 is weakly related to RL1. We do not rule out identification of a small number of additional virus-specific genes in future analyses.

Of the 189 unique genes originally proposed in AD169 by Chee *et al.* (1990), 108 remain unchanged as a result of subsequent reinterpretations and the present analysis (Fig. 3). We have discounted 46 as being unlikely to encode proteins, made minor revisions to 20, and identified five new AD169 genes (UL15A, UL21A, UL128, UL131A and US34A). AD169 RL13 and RL14 represent a frameshifted, and therefore non-functional, counterpart of a larger gene (RL13) in CCMV. We confirmed that this part of the AD169 sequence is correct and that the Toledo

gene is not frameshifted (data not shown), in accord with results published recently (Yu *et al.*, 2002). The ORF (IRL14) mapped at the right end of U_L by Chee *et al.* (1990) is spurious, as the 3' portion of UL148 is located here in a different reading frame. A frameshift is also present in a tract of eight T residues in the coding strand in the first exon of AD169 UL131A, and would render this gene non-functional. Again, we confirmed that the AD169 sequence is correct in this region. The corresponding exon in Toledo and in HCMV from the urine of a congenitally infected child was not frameshifted and contained a tract of seven T residues (data not shown).

The additional region at the right end of U_L in Toledo was interpreted by Cha *et al.* (1996) as containing 19 genes absent from AD169, in addition to UL130 and UL132 which are present in AD169. Using similar criteria to those used to compare CCMV and HCMV, we count a total of 23 genes, having redefined four, discounted five (UL134, UL137, UL143, UL149 and UL151), and introduced five novel genes in addition to UL131A and a disrupted form of UL128 (Fig. 3). This region of the Toledo genome is not collinear with the corresponding part of the CCMV genome. Inversion of a segment of the Toledo genome from a point immediately upstream of UL133 to a point between the second and third exons of UL128 would result in a collinear relationship consistent with the conclusions of Prichard *et al.* (2001), with UL148 adjacent to UL132 as in CCMV and AD169. These features indicate strongly that an inversion event has occurred during derivation of Toledo, and that this strain consequently lacks an intact UL128 gene. Our interpretation of genes at the right end of U_L in Towne is also shown in Fig. 3.

As far as can be ascertained, all CCMV genes are intact except UL128, which is frameshifted in the first exon. Re-examination of the corresponding region of the Colburn strain of simian cytomegalovirus (Chang *et al.*, 1995; accession no. U38308), which has been passaged many times in human fibroblasts (Huang *et al.*, 1978), showed the

Fig. 3. (on facing page) Layout of genes in the AD169 genome and the regions at the right end of U_L in the Toledo and Towne genomes that are absent from AD169. The scales are in kbp. In AD169, R_L and R_S are shown in a thicker format than U_L and U_S. Protein-coding regions are indicated by coloured arrows grouped according to the key, with gene nomenclature below. Introns are shown as narrow white bars. In AD169, genes in R_L and R_S are given their full nomenclature, but the UL and US prefixes have been omitted from UL1–UL148 (12–179 kbp) and US1–US34A (193–226 kbp). In Toledo, the UL prefix has been omitted. Colours indicate gene status in comparison with the layouts deduced by Chee *et al.* (1990) and Cha *et al.* (1996). Some ORFs are *unchanged*, and others have been modified or added as a result of *subsequent* studies (excepting the present study). Some have been *redefined* as a result of our study (by correction of a sequencing error, assignment or reassignment of an initiation codon or trivial correction of a coordinate). Some appear to be *non-functional* through frameshift mutation or partial deletion, and others are *novel* genes. UL131A is novel but non-functional in AD169 because of a frameshift mutation. UL128 is novel but non-functional in Toledo through inversion of part of the genome. The vertical line in the Toledo region downstream from UL148A indicates one end of the inverted region; the other end is immediately to the left of the region shown and places the third exon of UL128 upstream from UL133 and oriented leftwards. UL146, RL13 and RL12 in the Towne region were described as UL152, UL153 and UL154, respectively, in Cha *et al.* (1996). The locations of the 3' region of UL145 and the 5' region of UL1 (each leftward oriented) in this genome, immediately adjacent (as indicated by the vertical line) and in different reading frames, are presumably a result of replacement of genes at the right end of U_L by genes in TR_L and the adjacent part of the left end of U_L.

presence of a UL128 counterpart containing three exons as in HCMV and CCMV, with exon 2 frameshifted by loss of an A residue after nucleotide 1788. We confirmed this mutation in Colburn (data not shown). UL128 thus appears to be disrupted in CCMV, Toledo and Colburn, but intact in AD169.

The UL15A, UL147A and UL148D proteins contain a hydrophobic domain near their C termini, and the UL148A, UL148B and UL148C proteins contain a hydrophobic domain near their N termini. The sequences of the putative UL14 and UL141 glycoproteins are related, thus defining a new gene family (the UL14 family; Fig. 4). Similarly, conservation of an MHC-I domain in the UL18 (Beck & Barrell, 1988) and UL142 proteins defines the UL18 family (Fig. 5). The UL142 proteins are more diverged from each other and from MHC-I than are the UL18 proteins, and conservation of the MHC-I domain in the CCMV UL142 protein is greater than in the HCMV UL142 protein. Although Novotny *et al.* (2001) described structural motif predictions for the HCMV UL142 protein as unclear, they nonetheless noted its MHC-I-like nature.

We present evidence for an additional member (UL26) of the US22 gene family, in which Chee *et al.* (1990) numbered 12 genes. Members of this family are present in all sequenced members of the *Betaherpesvirinae*, and share one or more of four sequence motifs (I–IV; Nicholas, 1996). Recently, Stamminger *et al.* (2002) mentioned that

the UL26 protein contains a domain characteristic of US22 proteins (pfam02393). The alignments in Fig. 6 demonstrate the presence of motif III (part of pfam02393) in all previously identified HCMV US22 proteins and also in the UL26 protein. The UL26 protein is a component of the virion tegument (Baldick & Shenk, 1996; Stamminger *et al.*, 2002), as are other US22 proteins (Adair *et al.*, 2002), and contains a transcriptional activation domain (Stamminger *et al.*, 2002). Until now, US22 motifs have not been found outside the *Betaherpesvirinae*, but we have identified motif III in several other viruses from the *Herpesviridae*, *Poxviridae* and *Adenoviridae* (Fig. 6). Sequences more distantly related to motif III are also present in a subset of additional genes from the *Alphaherpesvirinae* (Fig. 6). No similar motif was found in non-viral proteins.

Three sizeable regions of the AD169 genome emerged as probably not encoding proteins, on the basis that they lack ORFs that are conserved in position or sequence. They are located at 2–8 kbp (in TR_L, and repeated in IR_L at 182–188 kbp), 91–99 kbp (containing the origin of lytic DNA replication) and 156–160 kbp (Fig. 3), in regions whose analysis for protein-coding potential was considered problematic by Chee *et al.* (1990). These regions display unusual features of local nucleotide composition (for example, regions of high A + T-content), and encode RNAs (e.g. the major early RNA encompassing RL4) with unknown functions that may not involve translation (Hutchinson *et al.*, 1986; Greenaway & Wilkinson, 1987; Plachter *et al.*,

HCMV UL14	MGGGRLPPLWLPL...LIAWSEWGNCC..LDAPPVVRSPCLQPVDRDRNRERNPGS
CCMV UL14	MAG..AIRLWLPLAAAIVAAVAPTGRCCDGCWCPPILRSPCLQMLSDRNRERRLSS
CCMV UL141	MARRQVGKVRRLPLLTVMGMLLEPEKSSRAGALLSDAAAPTGLPKETRAENYETVSPA
HCMV UL141	MCRRESLRTPWLFVWLLSCPRLLSEYSSSFPPFATADIAEKMW.....AENYETTSPA
HCMV UL14	PQLLPYGDRLEVACIFPAHDWPEVSIHVHLCYW..PEIVRSLVVDARSQVQLHNDASCYT
CCMV UL14	PKMLPYGDRLDVSCIFPSHAWPEVSIHVRLCYW..PEIVRSLVVDAHSGQLLHNDVTCQI
CCMV UL141	PMLVAEGERVTVPCVTMHIQWPLAVIRATFCLSPTPTERSVELVFNLSRGGEVVVPLEEKL
HCMV UL141	PVLVAEGEQVTIPCTVMTHSWPMVSIIRARECRS..HDGSDLEILDAVKGHRMLMGLQYRL
HCMV UL14	AGGR.....WRFEDGGAAQRLSLSFRLITETAGTYTCVLGNETHSLATETTALVADVHD
CCMV UL14	SSGR.....WSFEDGGAAQTLTSLFRLITQTAGTYHCVLGNETHGVVTESTALVADVHE
CCMV UL141	TESRLRGVSYNFTDLHVAHLFSLTFRATASSAGVYECILRNDSHALVLQRFVVLTHLET
HCMV UL141	PYAT.....WNFSQLHLGQIFSLTFNVSMDTAGMYECVLERNYSGLIMQRFVILTQLET
HCMV UL14	LRHSDR.SCDLAFGSRSTRYLWTPDPSRLRSINCGWEGEER..HRVVHYIPGTSGL....
CCMV UL14	LQKTDR.SCDLAFGSRSTQTKYLWTPDPSKLRSDCGWEGEER..HRLVHYIPGSSGE....
CCMV UL141	LSWADEPCCTPPLGRHSQDRQIWSPTPWRLKHVDCGWSSQMGFRNRNRYIEPASIGDGE
HCMV UL141	LSRDEPCCTPALGRYSLGDIWSPTPWRLRNHDCG..TYRGFQRNYFYI.....
HCMV UL14	LPSCEEDER.ELCVPFISQSIADNNCSRHRVDGARRR.....
CCMV UL14	EPSCEENEH.QLCVPIWKSADNNCSRQDREDEEEDDEPSSTAA
CCMV UL141	INIDAGEGEDEECWGPLCQPEEQPDRCWVPVVRQYRLPGGCYRSGDKQAKFVPALPTEPP
HCMV UL141	GRADAEDCWKPAC.PDEEPDRCWTVIQRYRLPGDCYRSQPHPPKFLPVT.....
HCMV UL14	YHLRRDYWLTDPKIGLLAAGSVALTSLCHLLCYWCSESYRRLNTEEESEAAEET
CCMV UL14	KRQPSSSQLSRDYWPSDPRIGVLAACSVVFTTTVHLLCYWCSKSYRRLTDEGEALYRRQ
CCMV UL141	LATPPEIPTGLSPWATRGVASFGLGFWGFFSVGVFCYLIYLCRRR.C....LGRR.GDGY
HCMV UL141	.APPADIDTGMSPWATRGTAAFLGFWISFTVCFLCYLCYLQCCGRWCPTPGRGRRGEGY
HCMV UL14	AAGEASAVAAAASVEEEQRRE
CCMV UL14	VAPETEESDRPGRDKGDEKSDPSSHLKDE
CCMV UL141	RRLPQYDEYPGHRMMKT
HCMV UL141	RRLPYDSYPGVRRMKR

Fig. 4. Amino acid sequence alignment of primary translation products of UL14 and UL141 in HCMV (Toledo) and CCMV. Putative transmembrane domains are underlined, and fully conserved residues are in bold.

		f	
	ooOoo	g oyOy	Oyrla sO xffr Gl y
HCMV UL23	91	SELYLGASGAMYLWTDHIYSDSLTFVAESITEFLNI	.GL....RRCNF
HCMV UL24	90	AVCLISDEGYVFCY..VREDTAVYYLARNLMEFARV	.GL..RAVELTH
HCMV UL28	>107	PEIWNVSGHGAFAY..LPGEDKVVYVLGLSFEFFEN	.GL....FAVY
HCMV UL29	290	RILCGDTGTVYAAL...VGQDKLVRLARDLRGFVRV	.GLALLIDDFRY
HCMV UL36	93	LIVLIGQRGGIYCY..DDLRCVYELAPTMKDFLRN	.GF...RHRDFH
HCMV UL43	340	ISILVNESGPVFGV..HPETROAHFLARGLLGFFRV	.GF..LRFCNNY
HCMV US22	241	GLVLLDKFGVVYLH..KIEDSDLYRIADNFHMLKC	.GLLKLRLGCRF
HCMV US23	280	VVIVKSHLDRSPPL..ORLAGEIYRLADSLEELFRA	.GLMKVYVRRRY
HCMV US24	239	VLVLLDWFGAVYAIQMDDPNHYVRRVANTITEFFRM	.GL....LKMVF
HCMV US26	137	LVLLLGRYETVWCL..DRDRGVLYYLAHSLDDFARH	.GL..LHCEAIY
HCMV TRS1/IRS1	461	LVVLLDELGAVFY...CPLDGHVYPLAAELSHFLRA	.GV..LGALALG
HCMV UL26	80	SVAGVAADGSVLCY.EISRENFVVRADSLPQLLER	.GL....LHSY
CCMV UL26	80	SVVGVGDDGSVLCY.EIGKENFVVRADSLAQLLER	.GL....LKS
RCMV R26	110	RVVAANGAGRICVY.DMGPVCFISRADDLVDFLAY	.GL....PDEY
TuHV-1 T26	116	KVTGIGDQAVYCYEFRAQNYVIRADDLETARQ	.GV....DLEY
RaHV-1 ORF93	83	PVLLADSGRIYLY...YKHKMWYAAYSLOHLL	.GV...CNLNY
FPV ORF250	47	LIILMSEDKIYVY...DDEALYKVADTMEEFSEI	.GL..INLGNVY
MDV ORF87	127	LLILVAEGGEVYAY...EEDTLHKLATSFSEFLEI	.GV..KSLGREVY
FAdV-10 ORF4	19	VVILVAEDKTVYGY...EEDTLHKLASTIPEFFRI	.GM.QNFGTEVF
BHV-1 circ	77	FMLYLMRGGAVYGY...EPTTGLHRLADSLHDFLT	.TAGL
EHV-1 ORF3	84	FLLYLSRNFQVFGY...EDGTGLHMLAKSLHDFLKF	.KGL
EHV-4 ORF3	84	FLLYLSRNFQVFGY...EDGTGLHMLAKTLHDFLKF	.KGL
MDV ORF71	72	FVILKGQDGNIYGV...ENASCIHFMSQNLHEFICK	.TGI
VZV ORF2	100	ILAMKEKRSTTLGY...ECGTGLHLLAPSMGTFLR	.THGL
HVT ORF66	72	FVIFEGQDGSVYGT...DNNVSLHFLGRNLHDILS	.KRG

Fig. 6. Conservation of motif III in members of the HCMV US22 gene family and similarity to other viral proteins. The location in the primary translation product of the first residue shown is indicated. The upper block of amino acid sequences shows an alignment of motif III in the 12 previously identified HCMV US22 proteins, with a consensus above (O, all residues hydrophobic; o, many residues hydrophobic; x, many residues acidic; upper-case character, all residues the appropriate residue; lower-case character, many residues the appropriate residue). The second block shows the corresponding region of the HCMV UL26 protein with its orthologues in other members of the *Betaherpesvirinae* (RCMV, AAF99126; TuHV-1, AAK57064). The third block shows potential examples of motif III in other virus proteins (RaHV-1, ranid herpesvirus 1, unpublished data; FPV, fowlpox virus, AAF44594; MDV, Marek's disease virus, AAG14261; FAdV-10, fowl adenovirus serotype 10, AAB88669). The fourth block shows sequences more distantly related to motif III in proteins from certain members of the *Alphaherpesvirinae* (BHV-1, bovine herpesvirus 1, AAB05201; EHV-1, AAB02438; EHV-4, AAC59545; MDV, AAG14251; VZV, varicella-zoster virus, CAA27885; HVT, herpesvirus of turkeys, AAG45796).

Fig. 7. Ka and Ks values (computed using GCG Diverge) for each homologous pair of CCMV and HCMV genes plotted against the position of the CCMV gene in its genome, with the CCMV genome arrangement (comprising major repeat and unique elements) depicted at the top of the diagram. For each pair of genes, the Ka and Ks values are marked by a filled circle and an open circle, respectively, and are joined by a solid line. For ten gene pairs where Ks was greater than 2.0, Ks is marked by a bar at 2.0. Nine gene pairs for which Ks values were not calculable are denoted at the top of the graph panel by filled squares. Median values for Ka (all gene pairs) and Ks (excluding the nine pairs with Ks greater than 2.0) are shown as dashed lines.

located nearer the genome termini (Fig. 2). Genes nearer the termini also exhibit generally higher levels of sequence divergence. This situation is reminiscent of the two sub-species of HHV-6 (Dominguez *et al.*, 1999).

Presently, we conclude that CCMV encodes 165 genes, each present in a single copy, with one (UL128) disrupted by a frameshift mutation. AD169 contains 145 genes, with four of these present in two copies in the R_L elements. Two AD169 genes (RL13 and UL131A) are disrupted, and a portion of UL148 (not counted in the total) is present at the right end of U_L . Revision of the coding potential of HCMV as described in Chee *et al.* (1990) and Cha *et al.* (1996) resulted in downgrading of 51 ORFs as probably not encoding proteins, minor corrections to 24, and the discovery of ten novel genes. Assuming that the wild-type HCMV genome approximates to the AD169 genome plus a rearrangement of the additional genes at the right end of U_L in Toledo, we infer a complement of 164 to 167 genes. The uncertainty results from the present inability to rule out the presence of CCMV UL155, UL156 and UL157 counterparts in HCMV, since the Toledo sequence in this region is unclear. Further refinement of the number and locations of genes in wild-type HCMV awaits the derivation of viral genome sequences directly from infected human tissue.

ACKNOWLEDGEMENTS

This work was supported by the Medical Research Council and research grant R01 AI24576 to G. S. H. from the National Institutes of Health, USA.

We are grateful to Richard Heberling (Esoterix Infectious Disease Center, 7540 Louis Pasteur, Suite 200, San Antonio, TX 78229, USA) for providing the CCMV isolate. We thank Lynne Neale and Gavin Wilkinson (University of Cardiff) for provision of DNA isolated from the urine of a child congenitally infected with HCMV, and Kathleen Wright for technical assistance with sequencing.

REFERENCES

- Adair, R., Douglas, E. R., Maclean, J. B., Graham, S. Y., Aitken, J. D., Jamieson, F. E. & Dargan, D. J. (2002). The products of human cytomegalovirus genes UL23, UL24, UL43 and US22 are tegument components. *J Gen Virol* **83**, 1315–1324.
- Bahr, U. & Darai, G. (2001). Analysis and characterization of the complete genome of tupaia (tree shrew) herpesvirus. *J Virol* **75**, 4854–4870.
- Baldick, C. J., Jr. & Shenk, T. (1996). Proteins associated with purified human cytomegalovirus particles. *J Virol* **70**, 6097–6105.
- Beck, S. & Barrell, B. G. (1988). Human cytomegalovirus encodes a glycoprotein homologous to MHC class I antigens. *Nature* **331**, 269–272.
- Cha, T. A., Tom, E., Kemble, G. W., Duke, G. M., Mocarski, E. S. & Spaete, R. R. (1996). Human cytomegalovirus clinical isolates carry at least 19 genes not found in laboratory strains. *J Virol* **70**, 78–83.
- Chambers, J., Angulo, A., Amaratunga, D. & 9 other authors (1999). DNA microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression. *J Virol* **73**, 5757–5766.
- Chan, Y. J., Chiou, C. J., Huang, Q. & Hayward, G. S. (1996). Synergistic interactions between overlapping binding sites for the serum response factor and ELK-1 proteins mediate both basal enhancement and phorbol ester responsiveness of primate cytomegalovirus major immediate-early promoters in monocyte and T-lymphocyte cell types. *J Virol* **70**, 8590–8605.
- Chang, Y., Jeang, K., Lietman, T. & Hayward, G. S. (1995). Structural organization of the spliced immediate-early gene complex that encodes the major acidic nuclear (ie1) and transactivator (ie2) proteins of African green monkey cytomegalovirus. *J Biomed Sci* **2**, 105–130.
- Chee, M. S., Bankier, A. T., Beck, S. & 12 other authors (1990). Analysis of the protein coding content of the sequence of human cytomegalovirus strain AD169. *Curr Top Microbiol Immunol* **154**, 125–169.
- Dargan, D. J., Jamieson, F. E., Maclean, J., Dolan, A., Addison, C. & McGeoch, D. J. (1997). The published DNA sequence of the human cytomegalovirus strain AD169 lacks 929 base pairs affecting genes UL42 and UL43. *J Virol* **71**, 9833–9836.
- Davis-Poynter, N. J., Lynch, D. M., Vally, H., Shellam, G. R., Rawlinson, W. D., Barrell, B. G. & Farrell, H. E. (1997). Identification and characterization of a G protein-coupled receptor homolog encoded by murine cytomegalovirus. *J Virol* **71**, 1521–1529.
- Davison, A. J. & Telford, E. A. R. (1994). Large scale DNA sequencing by manual methods. In *Methods Gene Technology*, vol. 2, pp. 151–175. Edited by J. W. Dale & P. G. Sanders. London: JAI Press.
- Dolan, A., Jamieson, F. E., Cunningham, C., Barnett, B. C. & McGeoch, D. J. (1998). The genome sequence of herpes simplex virus type 2. *J Virol* **72**, 2010–2021.
- Dominguez, G., Dambaugh, T. R., Stamey, F. R., Dewhurst, S., Inoue, N. & Pellett, P. E. (1999). Human herpesvirus 6B genome sequence: coding content and comparison with human herpesvirus 6A. *J Virol* **73**, 8040–8052.
- Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186–194.
- Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175–185.
- Goldmacher, V. S., Bartle, L. M., Skaletskaya, A. & 10 other authors (1999). A cytomegalovirus-encoded mitochondria-localized inhibitor of apoptosis structurally unrelated to Bcl-2. *Proc Natl Acad Sci U S A* **96**, 12536–12541.
- Gompels, U. A., Nicholas, J., Lawrence, G., Jones, M., Thomson, B. J., Martin, M. E. D., Efstathiou, S., Craxton, M. & Macaulay, H. A. (1995). The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution. *Virology* **209**, 29–51.
- Greenaway, P. J. & Wilkinson, G. W. (1987). Nucleotide sequence of the most abundantly transcribed early gene of human cytomegalovirus strain AD169. *Virus Res* **7**, 17–31.
- Huang, E. S., Kilpatrick, B., Lakeman, A. & Alford, C. A. (1978). Genetic analysis of a cytomegalovirus-like agent isolated from human brain. *J Virol* **26**, 718–723.
- Huang, L., Zhu, Y. & Anders, D. G. (1996). The variable 3' ends of a human cytomegalovirus oriLyt transcript (SRT) overlap an essential, conserved replicator element. *J Virol* **70**, 5272–5781.
- Hutchinson, N. I., Sondermeyer, R. T. & Tocci, M. J. (1986). Organization and expression of the major genes from the long inverted repeat of the human cytomegalovirus genome. *Virology* **155**, 160–171.

- Isegawa, Y., Mukai, T., Nakano, K. & 10 other authors (1999). Comparison of the complete DNA sequences of human herpesvirus 6 variants A and B. *J Virol* **73**, 8053–8063.
- Kotenko, S. V., Saccani, S., Izotova, L. S., Mirochnitchenko, O. V. & Pestka, S. (2000). Human cytomegalovirus harbors its own unique IL-10 homolog (cmvIL-10). *Proc Natl Acad Sci U S A* **97**, 1695–1700.
- Lockridge, K. M., Zhou, S. S., Kravitz, R. H., Johnson, J. L., Sawai, E. T., Blewett, E. L. & Barry, P. A. (2000). Primate cytomegaloviruses encode and express an IL-10-like protein. *Virology* **268**, 272–280.
- McGeoch, D. J., Dolan, A. & Ralph, A. C. (2000). Toward a comprehensive phylogeny for mammalian and avian herpesviruses. *J Virol* **74**, 10401–10406.
- Megaw, A. G., Rapaport, D., Avidor, B., Frenkel, N. & Davison, A. J. (1998). The DNA sequence of the RK strain of human herpesvirus 7. *Virology* **244**, 119–132.
- Mocarski, E. S. & Tan Courcelle, C. (2001). Cytomegaloviruses and their replication. In *Fields Virology*, 4th edn, vol. 2, pp. 2629–2673. Edited by D. M. Knipe & P. M. Howley. Philadelphia: Lippincott Williams & Wilkins.
- Mocarski, E. S., Prichard, M. N., Tan, C. S. & Brown, J. M. (1997). Reassessing the organization of the UL42-UL43 region of the human cytomegalovirus strain AD169 genome. *Virology* **239**, 169–175.
- Neote, K., DiGregorio, D., Mak, J. Y., Horuk, R. & Schall, T. J. (1993). Molecular cloning, functional expression, and signaling characteristics of a C-C chemokine receptor. *Cell* **72**, 415–425.
- Nicholas, J. (1996). Determination and analysis of the complete nucleotide sequence of human herpesvirus 7. *J Virol* **70**, 5975–5989.
- Novotny, J., Rigoutsos, I., Coleman, D. & Shenk, T. (2001). *In silico* structural and functional analysis of the human cytomegalovirus (HHV5) genome. *J Mol Biol* **310**, 1151–1166.
- Pass, R. F. (2001). Cytomegalovirus. In *Fields Virology*, 4th edn, vol. 2, pp. 2675–2705. Edited by D. M. Knipe & P. M. Howley. Philadelphia: Lippincott Williams & Wilkins.
- Penfold, M. E., Dairaghi, D. J., Duke, G. M., Saederup, N., Mocarski, E. S., Kemble, G. W. & Schall, T. J. (1999). Cytomegalovirus encodes a potent α chemokine. *Proc Natl Acad Sci U S A* **96**, 9839–9844.
- Plachter, B., Traupe, B., Albrecht, J. & Jahn, G. (1988). Abundant 5 kb RNA of human cytomegalovirus without a major translational reading frame. *J Gen Virol* **69**, 2251–2266.
- Prichard, M. N., Penfold, M. E. T., Duke, G. M., Spaete, R. R. & Kemble, G. W. (2001). A review of genetic differences between limited and extensively passaged human cytomegalovirus strains. *Rev Med Virol* **11**, 191–200.
- Rawlinson, W. D. & Barrell, B. G. (1993). Spliced transcripts of human cytomegalovirus. *J Virol* **67**, 5502–5513.
- Rawlinson, W. D., Farrell, H. E. & Barrell, B. G. (1996). Analysis of the complete DNA sequence of murine cytomegalovirus. *J Virol* **70**, 8833–8849.
- Smith, J. A. & Pari, G. S. (1995). Human cytomegalovirus UL102 gene. *J Virol* **69**, 1734–1740.
- Staden, R., Beal, K. F. & Bonfield, J. K. (2000). The Staden package, 1998. *Methods Mol Biol* **132**, 115–130.
- Stamminger, T., Gstaiger, M., Weinzierl, K., Lorz, K., Winkler, M. & Schaffner, W. (2002). Open reading frame UL26 of human cytomegalovirus encodes a novel tegument protein that contains a strong transcriptional activation domain. *J Virol* **76**, 4836–4847.
- Stenberg, R. M., Thomsen, D. R. & Stinski, M. F. (1984). Structural analysis of the major immediate early gene of human cytomegalovirus. *J Virol* **49**, 190–199.
- Stenberg, R. M., Witte, P. R. & Stinski, M. F. (1985). Multiple spliced and unspliced transcripts from human cytomegalovirus immediate-early region 2 and evidence for a common initiation site within immediate-early region 1. *J Virol* **56**, 665–675.
- Stenberg, R. M., Depto, A. S., Fortney, J. & Nelson, J. A. (1989). Regulated expression of early and late RNAs and proteins from the human cytomegalovirus immediate-early gene region. *J Virol* **63**, 2699–2708.
- Swinkels, B. W., Geelen, J. L., Wertheim-van Dillen, P., van Es, A. A. & van der Noordaa, J. (1984). Initial characterization of four cytomegalovirus strains isolated from chimpanzees. *Arch Virol* **82**, 125–128.
- Taylor, P. (1986). A computer program for translating DNA sequences into protein. *Nucleic Acids Res* **14**, 437–441.
- Telford, E. A. R., Watson, M. S., Perry, J., Cullinane, A. A. & Davison, A. J. (1998). The DNA sequence of equine herpesvirus-4. *J Gen Virol* **79**, 1197–1203.
- Tenney, D. J., Santomenna, L. D., Goudie, K. B. & Colberg-Poley, A. M. (1993). The human cytomegalovirus US3 immediate-early protein lacking the putative transmembrane domain regulates gene expression. *Nucleic Acids Res* **21**, 2931–2937.
- Vink, C., Beuken, E. & Bruggeman, C. A. (2000). Complete DNA sequence of the rat cytomegalovirus genome. *J Virol* **74**, 7656–7665.
- Yu, D., Smith, G. A., Enquist, L. W. & Shenk, T. (2002). Construction of a self-excisable bacterial artificial chromosome containing the human cytomegalovirus genome and mutagenesis of the diploid TRL/IRL13 gene. *J Virol* **76**, 2316–2328.