Journal of General Virology (2003), **84**, 657–663 DOI 10.1099/vir.0.18856-0

Short Communication

Homology between the human cytomegalovirus RL11 gene family and human adenovirus E3 genes

Andrew J. Davison,[1] Parvis Akter,[1] Charles Cunningham,[1] Aidan Dolan,[1] Clare Addison,[1] Derrick J. Dargan,[1] Aycan F. Hassan-Walker,[2] Vincent C. Emery,[2] Paul D. Griffiths[2] and Gavin W. G. Wilkinson[3]

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

[2]Department of Virology, Royal Free and University College Medical School, Royal Free Campus, Rowland Hill Street, Hampstead, London NW3 2QG, UK

[3]Section of Infection and Immunity, University of Wales College of Medicine, Tenovus Building, Heath Park, Cardiff CF14 4XX, UK

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

Received 25 September 2002
Accepted 6 December 2002

A significant proportion of the human cytomegalovirus (HCMV) genome comprises 12 multigene families that probably arose by gene duplication. One, the RL11 family, contains 12 members, most of which are predicted to encode membrane glycoproteins. Comparisons of sequences near the left end of the genome in several HCMV strains revealed two adjacent open reading frames that potentially encode related proteins: RL6, which is hypervariable, and RL5A, which has not been recognized previously. These genes potentially encode a domain that is the hallmark of proteins encoded by the RL11 family, and thus constitute two new members. A homologous domain is also present in a subset of human adenovirus E3 membrane glycoproteins. Evolution of genes specifying the shared domain in cytomegaloviruses and adenoviruses is characterized by extensive divergence, gene duplication and selective sequence loss. These features prompt speculation about the roles of these genes in the two virus families.

Human cytomegalovirus (HCMV; human herpesvirus 5) is the most complex human herpesvirus. The 229 354 bp genome of a high passage laboratory strain (AD169) was characterized by Chee *et al.* (1990) and predicted to contain 189 putative protein-coding genes, some duplicated in the inverted repeats. Subsequent revisions, based largely on comparisons with chimpanzee cytomegalovirus (CCMV), the closest known relative of HCMV, indicated that the wild-type HCMV genome contains 164–167 genes (Davison *et al.*, 2003).

Chee *et al.* (1990) identified nine multigene families in AD169, and an additional three were recognized in low passage isolates (Davison *et al.*, 2003). One, the RL11 family, consists of 12 genes (RL11, RL12, RL13, UL1, UL4, UL5, UL6, UL7, UL8, UL9, UL10 and UL11) oriented left to right near the left terminus of the genome and arranged contiguously but for the presence on the opposing strand of two unrelated genes (UL2 and UL3). RL11, RL12 and the 5′-portion of RL13 are present twice in the AD169 genome by virtue of their location in an inverted repeat. This repeat is much smaller in low passage strains, which consequently contain single copies of these genes (Prichard *et al.*, 2001). RL13 is also disrupted by a frameshift in AD169 (Yu *et al.*, 2002). All members of the RL11 family (generically termed 'RL11 genes') are dispensable for virus growth in cell culture (Ripalti & Mocarski, 1991; Hobom *et al.*, 2000; Yu *et al.*, 2002; Atalay *et al.*, 2002). RL11 genes are also present in CCMV, which has counterparts of all but UL1 (Davison *et al.*, 2003). Cytomegaloviruses of mouse, rat and tupaia lack RL11 genes (Rawlinson *et al.*, 1996; Vink *et al.*, 2000; Bahr & Darai, 2001).

The present work involved sequence analysis of the region near the left genome terminus in six HCMV strains. Four had been grown in human fibroblast cell lines, three of these (Merlin, 3157 and 6397) derived in Cardiff by three passages from urine samples from congenitally infected infants and one a widely used low passage strain (Toledo; Quinnan *et al.*, 1984). DNA from the Cardiff strains was obtained from purified virions and from Toledo as infected cell DNA. DNA was also prepared directly from clinical material for two strains, one (W) from the lung of an HCMV-infected AIDS patient and one (3301) from the urine of a congenitally infected child. Two overlapping fragments corresponding to nucleotides 928–5355 and 5052–9290 in the AD169 sequence were amplified by PCR from five of the DNA samples and cloned into pGEM-T (Promega). For each strain, the inserts in four plasmids were sequenced on both strands using universal and custom primers, and a consensus established. Corresponding data for the sixth

The GenBank accession numbers of the HCMV sequences reported in this paper are AY156040–AY156045.

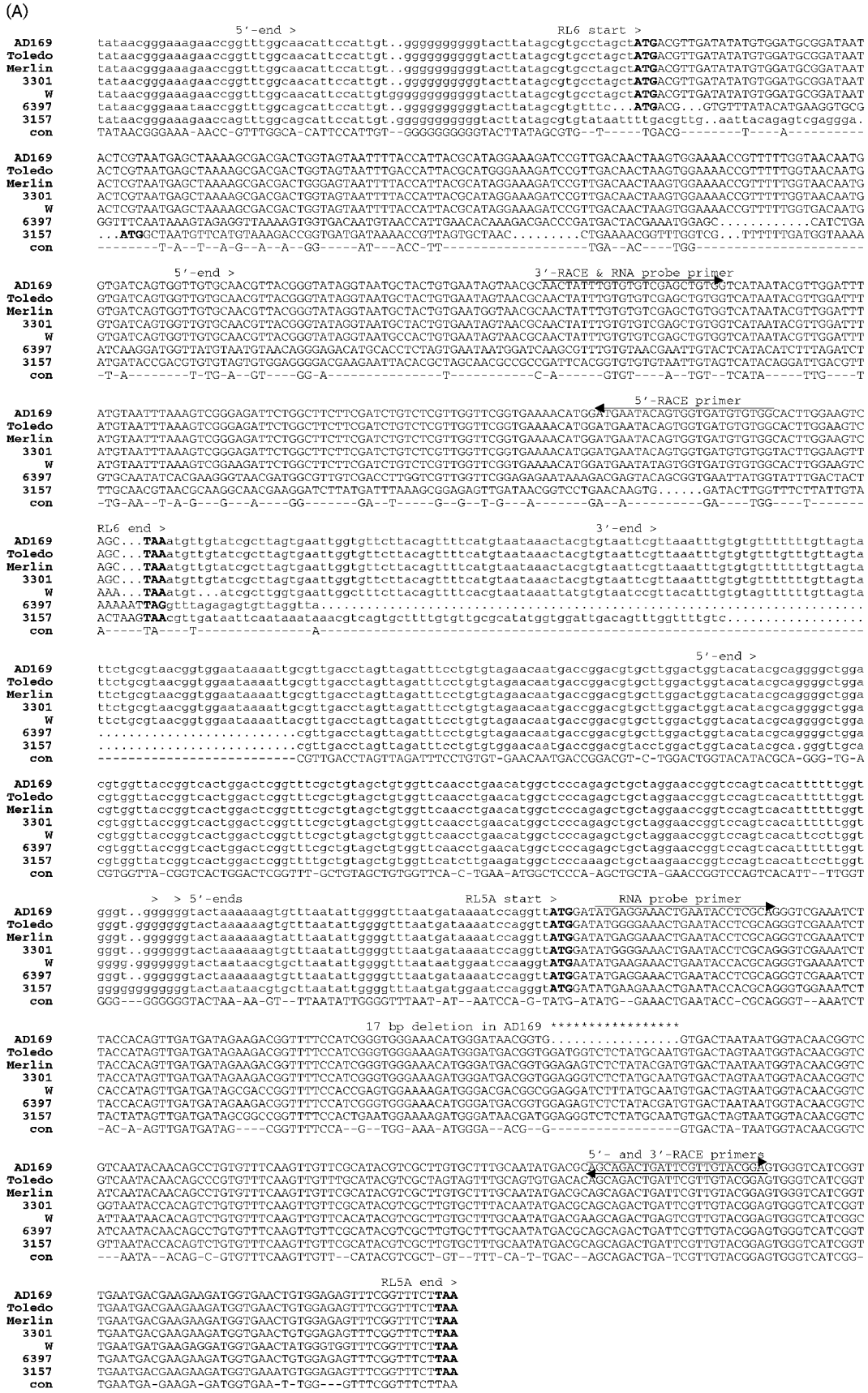

Fig. 1. For legend see page 659.

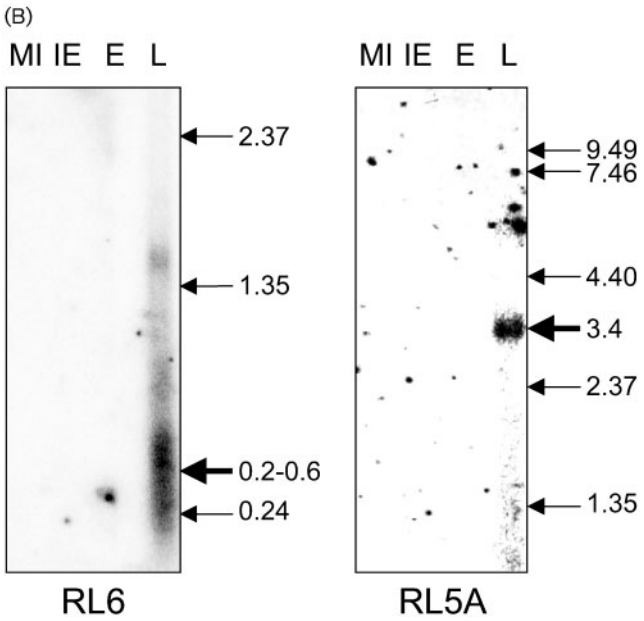

Fig. 1. (A) Alignment of the sequences of HCMV strains in the region containing RL6 and RL5A, with features of the AD169 sequence indicated. The orientation of the genes is right to left (i.e. opposite from the prototypic genome arrangement used in Fig. 3), with proposed protein-coding regions in upper-case and initiation and termination codons in bold. The genome coordinates of the AD169 sequence are 6015–4993 nucleotides. Where possible, the alignment was established by reference to amino acid sequence alignments. Conserved nucleotides are indicated in the 'con' line. With the exception of PCR artefacts (detected as infrequent nucleotide substitutions) and variable lengths of the two G tracts (which might also represent PCR artefacts), the sequences of the four plasmids determined for each strain were identical. Variability in the G tract preceding RL6 was noted for Merlin (9–10 residues), W (10–12), 6397 (9–11) and 3157 (10–12), and in the G tract preceding RL5A for W (9–14 residues) and 3157 (9–13). (B) Northern blots of polyadenylated RNA (3 μg per lane) extracted from mock-infected (MI) human foetal fibroblast cells or cells infected with HCMV (AD169) at an m.o.i. of 5. HCMV RNA was prepared under immediate-early (IE; 24 h in 200 μg cycloheximide ml^{-1}), early (E; 48 h in 300 μg phosphonoacetic acid ml^{-1}) and late (L; 72 h with no inhibitor) conditions. The probes were ^{32}P-labelled single-stranded RNAs. That for RL6 was derived from a PCR product extending from the RNA probe primer to the 5′-RACE primer in panel (A). That for RL5A was generated from a 5′-RACE product extending from the RNA probe primer to the 5′-RACE primer in panel (A). The positions of synthetic polyadenylated RNA markers (Life Technologies; visualized by ethidium bromide staining the gel before transfer, but not shown) are indicated, with sizes in kb [excluding the 3′-poly(A) tract]. The 0·2–0·6 kb RL6 and 3·4 kb RL5A mRNAs are shown by thicker arrows.

strain (Merlin) were obtained as part of shotgun M13 cloning the entire genome. Sequencing was carried out using ABI PRISM 377 and Beckman CEQ 2000XL instruments. Sequences were compiled using Pregap4 and Gap4 (Staden *et al.*, 2000) and Phred (Ewing & Green, 1998; Ewing *et al.*, 1998), and analysed using the GCG suite (Accelrys), Ptrans for sequence translation (Taylor, 1986), SignalP 2.0 for predicting signal cleavage peptides (Nielsen *et al.*, 1997) and the TMHMM2 module of the Smart suite for predicting transmembrane domains (Letunic *et al.*, 2002). The corresponding AD169 sequence was also included in the analysis.

An unusually variable region was noted in a small open reading frame (ORF), RL6. The protein encoded by a downstream ORF, RL5A, is related to the RL6 protein, but is less variable. An alignment of the region containing these two ORFs is shown in Fig. 1(A), and an alignment between the RL6 and RL5A amino acid sequences along the majority of their lengths is shown in the top section of Fig. 2. RL5A is disrupted by a frameshift mutation in AD169, and therefore was not identified by Chee *et al.* (1990). Moreover, HCMV RL6 was excluded along with many smaller ORFs described by Chee *et al.* (1990) as being unlikely to encode protein because of the lack of a counterpart in CCMV and the absence of data indicating functional expression (Davison *et al.*, 2003). Further comparisons confirmed that HCMV RL6 and RL5A are unique to HCMV.

The RL6 and RL5A proteins are related to each other and to RL11 proteins via a domain, termed here RL11D, and are thus new members of the RL11 family. The locations of RL11D in HCMV and CCMV proteins are illustrated in the top two sections of Fig. 3, and amino acid sequence alignments are compared in the top two sections of Fig. 2. RL11D is the key domain shared by the RL11 family, and corresponds essentially to an N-terminally extended form of a motif described by Chee *et al.* (1990) as $CXX(NQEKTY)X_{4–6}(YFLI)NX(ST)XXXXGXY$ (alternative residues in parentheses). RL11D consists of a region of variable length (65–82 residues) formed around three conserved residues (a tryptophan and two cysteines), with additional residues, including two potential *N*-linked glycosylation sites, often conserved near the extremities. The overall high level of variation is apparent from pairwise comparisons between any HCMV RL11 protein and its CCMV orthologue (Fig. 2).

Predictions for signal peptides and transmembrane domains in RL11 proteins are included in Fig. 3. The former are the more tentatively assigned. Some RL11 proteins contain a region rich in T and S residues located N-terminally (e.g. RL12) or C-terminally (e.g. UL11) to RL11D, indicative of *O*-glycosylation. Chee *et al.* (1990) noted that one RL11 protein (UL4) lacks a transmembrane domain and that two others (UL5 and UL8) have transmembrane domains but are N-terminally truncated, sharing similarity with RL11 proteins in regions outside RL11D. However, no evidence was obtained for splicing of UL4 to UL5 or UL8 (Chang *et al.*, 1989; Rawlinson & Barrell, 1993), and thus it appears that some RL11 proteins lack a signal peptide, RL11D or the transmembrane domain. The RL6 and RL5A proteins are in this class, lacking signal peptides and transmembrane

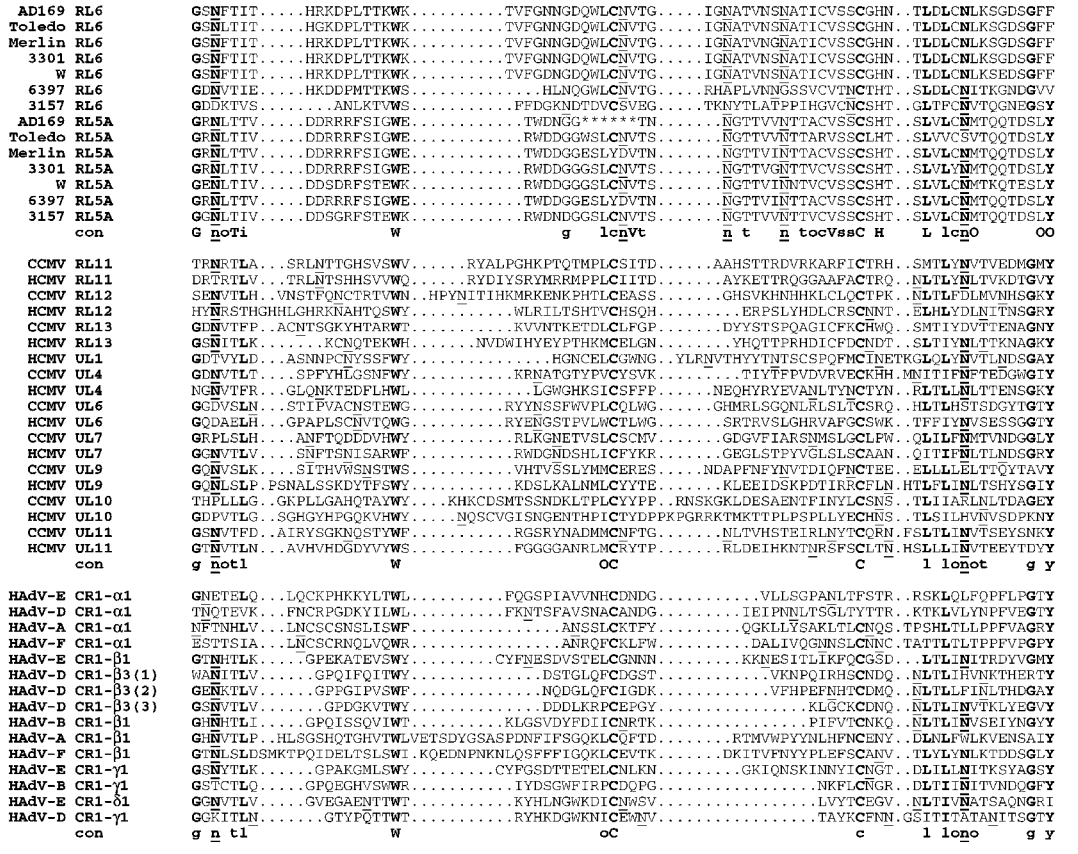

Fig. 2. Conservation of the RL11D domain in HCMV and CCMV RL11 proteins and the CR1 domain in HAdV E3 proteins. The upper section shows an alignment of RL11D in HCMV RL6 and RL5A amino acid sequences from several HCMV strains, with a consensus below (O, all residues hydrophobic; o, many residues hydrophobic; upper-case character, all residues the appropriate residue; lower-case character, many residues the appropriate residue). Asterisks indicate gapping characters required to join the frameshifted portions of AD169 RL5A. The second section shows an alignment of RL11D in previously recognized RL11 proteins, with a consensus below. The third section shows an alignment of CR1 in HAdV E3 proteins, with a consensus below. Residues fully (e.g. the tryptophan residue) or largely (e.g. the two cysteine residues) conserved in all RL11D and CR1 proteins are in bold. The line at the foot shows a consensus of 65 C-2 type Igs (smart00408 in the NCBI CDD), with more strongly conserved residues in bold and disulphide-bridging C residues doubly underlined. Tildes (~) indicate poorly conserved sequences of variable length. For each alignment, potential *N*-glycosylated residues are underlined.

domains, but retaining conserved potential *N*-linked glycosylation sites.

The ends of AD169 RL6 and RL5A late mRNAs shown in Fig. 1(A) were mapped using a Smart RACE kit (Clontech) with the primers indicated, followed by sequencing of cloned PCR products. 5′-ends for RL5A were identified in a tract of G residues and also about 120 bp upstream. The 3′ end was mapped to the right of the region shown in Fig. 1(A), at nucleotide 2090 in the AD169 sequence, downstream from a polyadenylation signal (AATAAA). Of the two 5′-ends mapped for RL6, the one upstream from the ORF is the better functional candidate, since the smaller RNA would direct translation of only 14 C-terminal residues. The 3′-end was located downstream from a polyadenylation signal. Northern blotting was carried out using strand-specific RNA probes prepared using Lig'nScribe and Maxiscript kits (Ambion), and identified late transcripts of 0·2–0·6 kb for RL6 and 3·4 kb for RL5A (Fig. 1B). These sizes are in accord with those deduced from RACE experiments (0·44 kb for RL6 and 3·4 kb for RL5A). The size range of RL6 transcripts may reflect the occurrence of two 5′-ends and the greater length heterogeneity of the 3′-poly(A) tract than in the synthetic markers. Minor RL6 transcripts were not investigated.

The arrangement of RL6 and RL5A mRNAs is illustrated in the upper section of Fig. 3. In accord with our results, Chambers *et al.* (1999) detected late transcription of RL6 using microarray technology. However, Hutchinson *et al.* (1986) identified neither mRNA by Northern blotting, but instead mapped a 2·0 kb early transcript on the other strand.

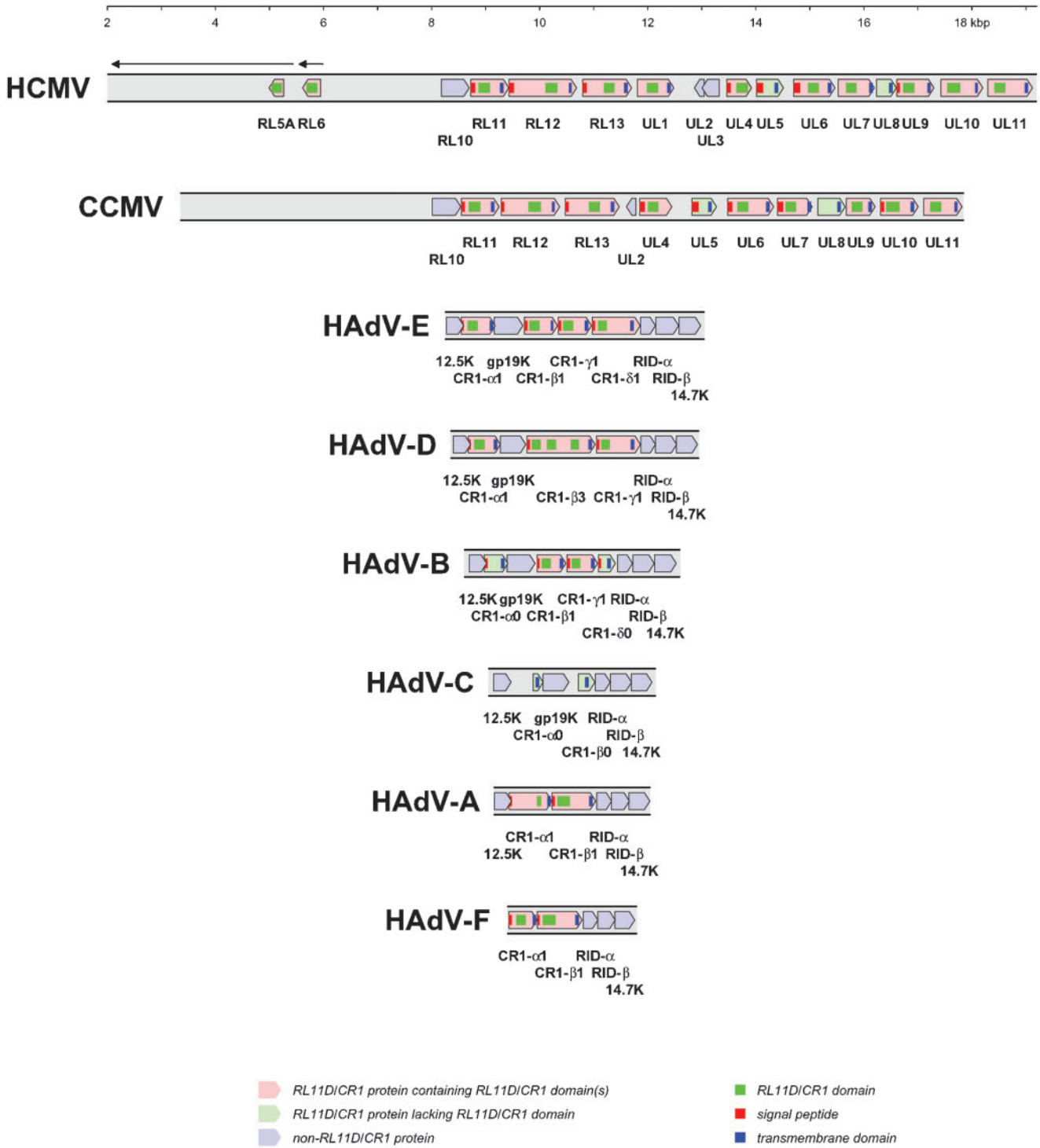

Fig. 3. Diagram showing the locations of RL11D or CR1 domains (considered equivalent), signal peptides and transmembrane domains in HCMV (AD169) and CCMV RL11 proteins and HAdV E3 proteins. Signal peptides and transmembrane domains are not indicated for proteins not considered to be in the RL11 or CR1 families. The scale shows coordinates (kbp) relative to the left end of the prototypic AD169 genome. RL13 and RL5A are frameshifted in AD169, but each is shown as a single ORF. mRNAs for HCMV RL6 and RL5A are depicted by horizontal arrows. Interpretations of gene content are shown for the cytomegaloviruses and for a single serotype in each HAdV species (H and S prefixes indicating human and chimpanzee viruses, respectively): HCMV (accession X17403 from nucleotide 2001) and CCMV (AF480884 from 2001) HAdV-12 for HAdV-A (X73487 from 26301), SAdV-21 for HAdV-B (AR101858 from 27401), HAdV-2 for HAdV-C (J01917 from 27801), HAdV-17 for HAdV-D (AF108105 from 26101), SAdV-25 for HAdV-E (AR101859 from 27101) and HAdV-40 for HAdV-F (L19443 from 26201).

This implies that RL6 and RL5A may be transcribed at relatively low levels. Direct evidence for protein expression is not available, but the transcript mapping data, sequence similarities of RL6 and RL5A to RL11 genes and high variability of RL6 (which would not be expected of a non-coding region) support the conclusion that the two genes are translated into small RL11D proteins.

We also discovered that an RL11D-like domain is present in proteins encoded by the E3 region of human adenoviruses (HAdVs), which is dispensable for virus growth in cell culture. The lower sections in Fig. 3 show the gene layout in the E3 region of the six HAdV species, developed from the scheme of Windheim & Burgert (2002). The three genes on the right are conserved, and encode two membrane proteins (RID-α and RID-β) and one non-membrane-associated protein (14·7K), which are involved in inhibition of apoptosis (reviewed by Russell, 2000). Except in HAdV-F, the left gene encodes a non-membrane-associated protein of unknown function (12·5K). Located between are two to five genes encoding membrane proteins, some of which are related via a domain defined in three HAdV proteins as CR1 (Deryckere & Burgert, 1996; pfam02440 in the NCBI CDD). Our examination of the data led to a redefinition of CR1 (removal of 5 residues from the N terminus and 17 from the C terminus, with an alternative alignment) and thence to the identification of several additional HAdV CR1 proteins, as shown in Figs 2 and 3. Additional analyses confirmed that some E3 proteins lack CR1 but are in other respects related to proteins that possess it. Thus, the nomenclature in Fig. 2 recognizes whether a gene contains (or its ancestor is proposed to have contained) CR1, the relative location of the gene (α, β, γ or δ), and the number of CR1 domains in the protein (0, 1 or 3). It is possible that an ancestor of gp19K was a CR1 protein, but no evidence remains in extant genomes. Therefore, the nomenclature has not been extended to this glycoprotein, which is involved in inhibition of the immune response by binding to MHC-I. Some CR1 proteins contain a region rich in T and S residues located C-terminally to CR1, indicative of *O*-glycosylation. The homology between adenovirus CR1 and cytomegalovirus RL11D is illustrated by the alignments in Fig. 2.

We conclude that certain cytomegalovirus RL11 and adenovirus E3 proteins share a domain which we term CR1 for both virus families in the following discussion (i.e. RL11D and CR1 are equivalent). The canonical CR1 protein is a class I membrane protein that contains CR1 in the external portion and is *N*- and perhaps *O*-glycosylated. Both forms of glycosylation have been confirmed experimentally in HAdV-B CR1-γ1 (Hawkins & Wold, 1995) and HAdV-D CR1-β3 (Windheim & Burgert, 2002), and HCMV UL4 has been characterized as a virion-associated *N*-glycosylated protein (Chang *et al.*, 1989). The wide evolution of CR1 genes is apparent in Fig. 2 from the overall poor sequence conservation and in Fig. 3 from the variable number of such genes in related virus species. Hypervariability has been documented for some RL11 genes in regions encompassing CR1, including RL11 and UL11 (Cha *et al.*, 1996; Hitomi *et al.*, 1997), and is now apparent in RL6. Moreover, both virus families encode truncated proteins that appear to have originated from CR1 proteins but lack CR1 (e.g. HCMV UL5 and UL8, HAdV-B CR1-α0 and CR1-δ0). These have presumably been selected for functions that do not depend on CR1. Cytomegaloviruses also encode CR1 proteins that may not be intrinsically membrane-associated (e.g. HCMV UL4, RL6 and RL5A).

CR1 proteins have been documented in cytomegaloviruses and adenoviruses of humans and chimpanzees, and possibly function as modulators of a family of variable host proteins. Similarities of CR1 to certain of the immunoglobulin domain (IgD) family (Halaby *et al.*, 1999) have been registered (Chee *et al.*, 1990; Windheim & Burgert, 2002), and a similarity between adenovirus CR1 and IgD is implicit in the NCBI CDD. Lilley *et al.* (2001) and Atalay *et al.* (2002) showed that HCMV RL11 encodes an Ig-binding protein, but the sequence alignment given in the latter publication in support of an IgD is unconvincing. Nonetheless, Fig. 2 hints at a relationship between CR1 and IgD (exemplified by C2 Igs) in regions near the extremities. The two disulphide-linked cysteine residues characteristic of certain IgD-containing proteins are not present in CR1, but the two cysteine residues conserved in the central region may form an alternative bridge. Thus, it remains possible that CR1 is a member of the IgD family.

ACKNOWLEDGEMENTS

This work was supported by the Medical Research Council, the Wellcome Trust and the Royal Society. We thank Cardiff Public Health Laboratory and Lynne Neale for primary culture of HCMV isolates and provision of strain 3301 DNA, and Brian McSharry for supplying purified strain Merlin DNA.

REFERENCES

Atalay, R., Zimmermann, A., Wagner, M., Borst, E., Benz, C., Messerle, M. & Hengel, H. (2002). Identification and expression of human cytomegalovirus transcription units coding for two distinct Fcγ receptor homologs. *J Virol* **76**, 8596–8608.

Bahr, U. & Darai, G. (2001). Analysis and characterization of the complete genome of tupaia (tree shrew) herpesvirus. *J Virol* **75**, 4854–4870.

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.

Chang, C.-P., Vesole, D. H., Nelson, J., Oldstone, M. B. A. & Stinski, M. F. (1989). Identification and expression of a human cytomegalovirus early glycoprotein. *J Virol* **63**, 3330–3337.

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.

Davison, A. J., Dolan, A., Akter, P., Addison, C., Dargan, D. J., Alcendor, D. J., McGeoch, D. J. & Hayward, G. S. (2003). The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome. *J Gen Virol* **84**, 17–28.

Deryckere, F. & Burgert, H. G. (1996). Early region 3 of adenovirus type 19 (subgroup D) encodes an HLA-binding protein distinct from that of subgroups B and C. *J Virol* **70**, 2832–2841.

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.

Halaby, D. M., Poupon, A. & Mornon, J.-P. (1999). The immunoglobulin fold family: sequence analysis and 3D structure comparison. *Protein Eng* **12**, 563–571.

Hawkins, L. K. & Wold, W. S. (1995). The E3–20.5K membrane protein of subgroup B human adenoviruses contains *O*-linked and complex *N*-linked oligosaccharides. *Virology* **210**, 335–344.

Hitomi, S., Kozuka-Hata, H., Chen, Z., Sugano, S., Yamaguchi, N. & Watanabe, S. (1997). Human cytomegalovirus open reading frame UL11 encodes a highly polymorphic protein expressed on the infected cell surface. *Arch Virol* **142**, 1407–1427.

Hobom, U., Brune, W., Messerle, M., Hahn, G. & Koszinowski, U. H. (2000). Fast screening procedures for random transposon libraries of cloned herpesvirus genomes: mutational analysis of human cytomegalovirus envelope glycoprotein genes. *J Virol* **74**, 7720–7729.

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.

Letunic, I., Goodstadt, L., Dickens, N. J. & 7 other authors (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res* **30**, 242–244.

Lilley, B. N., Ploegh, H. L. & Tirabassi, R. S. (2001). Human cytomegalovirus open reading frame TRL11/IRL11 encodes an immunoglobulin G Fc-binding protein. *J Virol* **75**, 11218–11221.

Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10**, 1–6.

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.

Quinnan, G. V., Jr, Delery, M., Rook, A. H. and others (1984). Comparative virulence and immunogenicity of the Towne strain and a nonattenuated strain of cytomegalovirus. *Ann Intern Med* **101**, 478–483.

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.

Ripalti, A. & Mocarski, E. S. (1991). The products of human cytomegalovirus genes UL1–UL7, including gp48, are dispensable for growth in cell culture. In *Prog Cytomegalovirus Research: Proc Third International Cytomegalovirus Workshop*, pp. 57–62. Edited by M. P. Landini. Amsterdam: Elsevier.

Russell, W. C. (2000). Update on adenovirus and its vectors. *J Gen Virol* **81**, 2573–2604.

Staden, R., Beal, K. F. & Bonfield, J. K. (2000). The Staden package, 1998. *Methods Mol Biol* **132**, 115–130.

Taylor, P. (1986). A computer program for translating DNA sequences into protein. *Nucleic Acids Res* **14**, 437–441.

Vink, C., Beuken, E. & Bruggeman, C. A. (2000). Complete DNA sequence of the rat cytomegalovirus genome. *J Virol* **74**, 7656–7665.

Windheim, M. & Burgert, H. G. (2002). Characterization of E3/49K, a novel, highly glycosylated E3 protein of the epidemic keratoconjunctivitis-causing adenovirus type 19a. *J Virol* **76**, 755–766.

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.