Bradley, Amanda J. (2008) *Analysis of two hypervariable human cytomegalovirus genes, UL146 and UL139.* PhD thesis.

http://theses.gla.ac.uk/432/

Analysis of two hypervariable human cytomegalovirus genes, UL146 and UL139

Amanda Jane Bradley

2008

A thesis presented to the Faculty of Biomedical and Life Science at the University of Glasgow for the degree of Doctor of Philosophy

MRC Virology Unit
Institute of Virology
Church Street
Glasgow G11 5JR

Abstract

Human cytomegalovirus (HCMV) is a highly host-specific, ubiquitous herpesvirus that results in asymptomatic infection for the majority of those infected. However, it produces serious clinical disease in neonates and immunocompromised individuals such as transplant recipients and AIDS patients. The majority of the 236 kbp genome is highly conserved, but there are a number of highly variable regions, coding and non-coding, scattered throughout the genome. Numerous studies have been published investigating the genotypes of hypervariable genes, most focussed on potential associations between genotype and clinical disease or tropism. In general, no convincing connections between genotype and disease have been found.

The present study investigated two hypervariable HCMV genes, UL146 and UL139, in a large number of clinical samples (179) from a number of locations worldwide in Europe, Africa, Asia and Australia. A total of 14 UL146 genotypes (G1-G14) were detected, which agrees with previous findings based on many fewer samples. For UL139, eight genotypes were detected, three of them (G5, G7 and G8) novel. The genotypes of both genes appear to have evolved under constraint rather than positive selection. Possible bias in the geographical distribution of the UL146 and UL139 genotypes was investigated. In general, all genotypes were found in all areas and any variation from the expected distribution was probably a result of small sample numbers from certain regions, specifically Asia and Australia. This general finding is in agreement with that of a previously published study on gene UL73.

No evidence for linkage disequilibrium between UL146 and UL139 genotypes was found. This is in accordance with a previously published study of linkage disequilibrium among six other genes (UL55, UL74, UL75, UL115, US9 and US28), and is consistent with the theory that recombination has played a role in HCMV evolution. The absence of linkage between highly variable genes complicates attempts to examine associations between genotype and disease, as many combinations of genotypes are possible.

Investigation of transcriptional expression of UL146 and UL139 from HCMV strain Merlin in fibroblast cell culture revealed that UL146 is expressed with late

kinetics and UL139 with early-late kinetics. Northern blot and RACE data suggested that UL146 is 3'-coterminally expressed with UL147, UL147A, UL148 and UL132, and that UL139 is 3'-coterminally expressed with UL140 and UL141.

To determine whether the high degree of sequence divergence corresponds to structural divergence, the UL146 genotypes were homology modelled on the related human chemokines IL-8, gro-α and IF9S. All 14 genotypes were predicted to be structurally very similar, which suggests they may also be functionally similar. However, small differences between the structures of human chemokines are known to result in slightly differing binding affinities for cellular receptors, and therefore even small differences between UL146 genotypes could conceivably confer functional differences.

UL139 has been predicted to encode a type 1 membrane glycoprotein. No information has been published regarding UL139 function, although a short region of similarity with the cellular signal transducer CD24 has been noted previously, tentatively suggesting an immunomodulatory role. Preliminary experiments to characterise UL139 were performed utilising recombinant adenovirus vectors expressing tagged UL139 variants from three genotypes (G1, G5 and G7). The tagged UL139 variants expressed proteins that were considerably larger in mass than predicted from amino acid sequences. This extra mass may be attributable to glycosylation as well as other forms of post-translational modification.

Mixed infections of HCMV strains in immunocompromised individuals, such as transplant recipients, have been associated with enhanced pathogenesis and increased risk of transplant rejection. The presence of mixed infections also further complicates attempts to establish connections between genotype and disease outcome. In the analysis of UL146 and UL139 genotypes, multiple genotypes were detected in 14% of samples and in 29% when repeated experimental results were included, and even these values may be underestimations. The utility of a QPCR-based assay using genotype-specific primers was assessed as a means of more accurately determining the occurrence of mixed infections, and showed promise.

Passage of HCMV strains in cell culture has been shown to result in various mutations. AD169, a commonly used laboratory strain, lacks 15 kbp sequence that includes UL146 and UL139. An alternative stock of AD169 (AD169*var*UC) was obtained that was thought to contain most or all of the deleted region and, indeed, both UL146 and UL139 were detected. Further sequencing confirmed that this stock is derived from AD169 and revealed that it contains all but 3.2 kbp of the 15 kbp absent from commonly used AD169 stocks. The 3.2 kbp deletion affects UL144, UL142, UL141 and UL140. This propensity of HCMV to undergo mutation during cell culture highlights the importance of studying characterised strains that are as close to wild type virus as possible.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Firstly I would like to thank my supervisor, Dr Andrew Davison for his support, advice and encouragement over the last four years. His patience during the preparation of this thesis was greatly appreciated, as were the diagrams he kindly provided. I would also like to thank Duncan for allowing me to do a PhD at the institute and his good humour. Thanks to everyone in lab 200, especially Charles, for their help, support, chat and laughter. I’d also like to thank Derrick Dargan and the rest of lab 204 for all their help. A big thank you to Derek Gatherer for his endless patience and advice with the bioinformatic and statistical analysis as well as the homology modelling.

A special thank you to all the collaborators who kindly donated samples for use in the genotyping study and Ida for her hard work on the Hungarian samples. In addition, I’d like to thank Gavin Wilkinson and his group in Cardiff for the use of their adenovirus system as well as their useful advice and scientific know-how.

I’d like to thank all the friends I’ve made here, Tanya, Sarah, Jon, James, Ali, Lorraine, Karl, David, Ross, Martin and Louise. Thanks for the coffee and chat, those fabulous parties and understanding how difficult this whole process is.

Thanks to Mam, for your constant support and endless faith in my ability. Finally I’d like to thank Rob for your patience, love and support.

Authors Declaration

The author was a recipient of a Medical Research Council Studentship. Except where specified, all of the results described in this thesis were obtained by the author’s own efforts.

Amanda Bradley

Abbreviations

aa	amino acid residues
AIDS	acquired immune deficiency syndrome
BAC	bacterial artificial chromosome
bp	base pairs
˚C	degrees Celsius
CCMV	chimpanzee cytomegalovirus
cDNA	complementary deoxyribonucleic acid
CMV	cytomegalovirus
CPE	cytopathic effect
DMSO	dimethylsulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleoside triphosphate
dsDNA	double-stranded deoxyribonucleic acid
DTT	dithiothreitol
E	early
EBV	Epstein-Barr virus
EDTA	ethylenediamine tetra-acetic acid
EHV-2	Equine herpesvirus 2
E-L	early-late
EM	electron micrograph
FCS	foetal calf serum
g	grams
gp	glycoprotein
h	hours
HCMV	human cytomegalovirus
HFFF	human foetal foreskin fibroblast
HHV-6	human herpesvirus 6
HHV-7	human herpesvirus 7
HHV-8	human herpesvirus 8
HSV-1	herpes simplex virus type 1
HSV-2	herpes simplex virus type 2
IE	immediate-early
IPTG	isopropyl-β-D-thiogalactoside
IR_L	internal long repeat
IR_S	internal short repeat
Ka	ratio of non-synonymous substitutions per possible non-synonymous site
Ks	ratio of synonymous substitutions per possible synonymous site
kb	kilobases
kbp	kilobase pairs
KSHV	kaposi’s sarcoma-associated herpesvirus
L	late
LD	linkage disequilibrium
M	molar
MCMV	murine cytomegalovirus
MHC	major histocompatibility complex
MI	mock infected
MIE	major immediate early
mRNA	messenger ribonucleic acid
mg	milligrams

min	minutes
ml	millilitres
mM	millimolar
m.o.i.	multiplicity of infection
MOPS	[3-(N-morpholino) propanesulfonic acid]
μg	micrograms
μl	microlitres
μM	micromolar
nm	nanometres
nt	nucleotides
ORF	open reading frame
PAA	phosphonoacetic acid
PCR	polymerase chain reaction
pfu	plaque forming units
polyA	polyadenylated
p.i.	post infection
RACE	rapid amplification of cDNA ends
RCMV	rat cytomegalovirus
RhCMV	rhesus cytomegalovirus
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
RT	reverse transcriptase
s	seconds
SCMV	simian (African green monkey) cytomegalovirus
SDS	sodium dodecyl sulphate
SSC	standard saline citrate
ssRNA	single-stranded ribonucleic acid
STR	short tandem repeat
TBE	Tris-borate with EDTA
TR_L	long terminal repeat
TR_S	short terminal repeat
TuHV-1	tupaiid herpesvirus 1
U_L	unique long
UPM	universal primer mix
U_S	unique short
UV	ultraviolet
VZV	varicella-zoster virus
v/v	volume/volume
WGA	whole genome amplification
w/v	weight/volume

1 Introduction

1.1 The family *Herpesviridae*

The family *Herpesviridae* consists of large, double-stranded DNA (dsDNA) viruses that have a distinctive virion structure, in which the genome is packaged in an icosahedral capsid surrounded by a proteinaceous tegument layer within a host-derived envelope decorated with viral glycoproteins. The human cytomegalovirus (HCMV) virion will be described in detail in Section 1.3.4. Herpesviruses have been isolated in a wide variety of hosts from mammals to invertebrates, and over 200 have been identified to date. All characterised herpesviruses have been shown to establish and maintain latent infections in their natural host. Further studies have shown that viral replication and capsid assembly occur in the nucleus, whereas maturation takes place in the cytoplasm. Production of infectious virus particles inevitably results in cell death.

The family *Herpesviridae* contains three subfamilies, the *Alphaherpesvirinae*, the *Betaherpesvirinae* and the *Gammaherpesvirinae* (Davison *et al.,* 2005a). The *Alphaherpesvirinae* are neurotropic and establish latency in neuronal ganglia. They exhibit broad host species range *in vitro*. Members that infect humans include herpes simplex virus types 1 and 2 and varicella-zoster virus (HSV-1, HSV-2 and VZV, respectively). The *Betaherpesvirinae* are characterised by slow replication in cell culture and restricted host range. They establish latency in peripheral blood monocytes (Kondo *et al.,* 1991; Kondo *et al.,* 1994). Members include cytomegaloviruses (CMVs) such as HCMV (also known as human herpesvirus 5 (HHV-5)) and murine cytomegalovirus (MCMV), and also human herpesviruses 6 and 7 (HHV-6 and HHV-7). The *Gammaherpesvirinae* are lymphotropic and infect lymphocytes *in vitro*. They establish latency in lymphocytes. Members include Epstein-Barr virus (EBV), equine herpesvirus 2 (EHV-2) and Kaposi’s sarcoma-associated herpesvirus (KSHV) (which is also known as human herpesvirus 8 (HHV-8)).

Recently, two new herpesvirus families have been established, the *Alloherpesviridae* (fish and frog herpesviruses) and *Malacoherpesviridae* (containing a bivalve herpesvirus) classified with the revised *Herpesviridae* (mammal, bird and reptile herpesviruses) into an order, the *Herpesvirales* (see www.ictvnet.org) (Davison, 2002, 2002a, 2002b; Davison *et al.*, 2005; McGeoch *et al.*, 2006). Table 1.1 compares the previous and revised classification schemes.

Table 1.1. Previous and Revised Herpesvirus Classification

Taxon Level	*Previous Taxon*	*Revised Taxon*	*Examples*
Order		*Herpesvirales*	
Family	*Herpesviridae*	*Herpesviridae*	
Subfamily	*Alphaherpesvirinae*	*Alphaherpesvirinae*	
Genus	*Simplexvirus*	*Simplexvirus*	HSV-1 HSV-2
Genus	*Varicellovirus*	*Varicellovirus*	VZV
Genus	*Mardivirus*	*Mardivirus*	Marek's disease virus
Genus	*Iltovirus*	*Iltovirus*	Infectious laryngotracheitis virus
Subfamily	*Betaherpesvirinae*	*Betaherpesvirinae*	
Genus	*Cytomegalovirus*	*Cytomegalovirus*	HCMV, chimpanzee CMV, rhesus CMV, African green monkey CMV
Genus	*Muromegalovirus*	*Muromegalovirus*	MCMV, rat CMV
Genus	*Roseolovirus*	*Roseolovirus*	HHV-6, HHV-7
Genus		*Proboscivirus*	Endotheliotropic elephant herpesvirus
Subfamily	*Gammaherpesvirinae*	*Gammaherpesvirinae*	
Genus	*Lymphocryptovirus*	*Lymphocryptovirus*	EBV
Genus	*Rhadinovirus*	*Rhadinovirus*	Herpesvirus saimiri, KSHV
Genus		*Macavirus*	Malignant catarrhal fever virus
Genus		*Percavirus*	EHV-2
Family		*Alloherpesviridae*	
Genus	*Ictalurivirus*	*Ictalurivirus*	Channel catfish virus
Family		*Malacoherpesviridae*	
Genus		*Ostreavirus*	Oyster herpesvirus

Herpesvirus genomes vary considerably in size, the smallest being that of simian varicella virus (SVV) at 124 kbp and the largest being that of koi herpesvirus at 295 kbp (Aoki *et al.*, 2007; Gray *et al.*, 2001). Their G+C content also varies greatly, ranging from 32-75 % (Honess, 1984). All members of the *Alpha-*, *Beta-*, and *Gammaherpesvirinae* are characterised by a set of 43 'core' genes that have been inherited from a common ancestor, although one or two of these genes have been lost in some lineages. Most of the core genes encode proteins essential for viral DNA replication, viral DNA packaging and capsid structure and assembly. Herpesviruses have evolved through nucleotide substitution, gene capture, gene duplication, recombination and genetic rearrangements. Other genes appear to be conserved only within subfamilies, genera or species (Davison *et al.,* 2003, 2003a).

Phylogenetic analyses of sequences from host species and the herpesviruses that infect them often reveal similar branching patterns, which suggests a large degree of co-evolution between virus and host (McGeoch *et al.,* 2000; McGeoch *et al.,* 2005). Figure 1.1 shows the five classes of herpesvirus genome structure characterised adequately to date. All contain unique regions bounded by internal or terminal repeat sequences in direct or inverse orientations. An example of class A is HHV-6, B, KSHV; C, EBV; D, VZV; E, HSV-1, HSV-2 and HCMV.

1.2 Human herpesviruses

Eight human herpesviruses have been discovered to date representing all three subfamilies of the family *Herpesviridae.* Many infections by human herpesviruses are asymptomatic, but they can result in severe or even fatal disease in the very young or immunocompromised. Three members of the *Alphaherpesvirinae* infect humans: HSV-1, HSV-2 and VZV. HSV-1 and HSV-2 are closely related and cause clinically similar diseases, although HSV-1 is more commonly associated with oral lesions whereas HSV-2 is associated with genital lesions (Efstathiou and Preston, 2005). VZV causes chickenpox, a rash of small vesicles that rupture and can cause intense itching, and reactivation causes zoster (shingles) (Gershon *et al.,* 1997).

Three members of the *Betaherpesvirinae* infect humans: HCMV, HHV-6 and HHV-7. HCMV is the major infectious cause of congenital disease (Gandhi and Khanna, 2004). HHV-6 is related to HCMV and 67% of HHV-6 proteins are homologous to proteins in HCMV U_L. However, the HHV-6 genome organisation differs significantly from that of HCMV (Figure 1.1) as it is composed of a unique long (U_L) region bounded by terminal direct repeats (Gompels *et al.,* 1995). HHV-6 has two variants, HHV-6A and HHV-6B. HHV-6B is more commonly associated with exanthem subitum, a febrile illness in children, and occasionally febrile seizures (Dewhurst *et al.,* 1997; Kosuge, 2000).

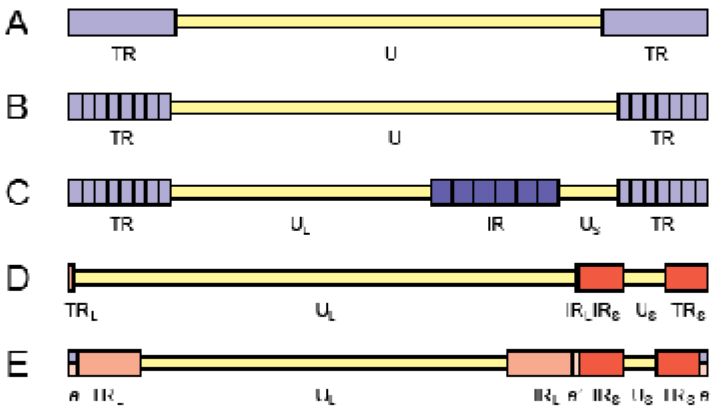


Figure 1.1 Genome structures of herpesviruses

Unique sequences are represented in yellow, direct repeat elements in blue and inverted repeats in red (not to scale). Repeats in different shades of the same colour are not related to each other. Class E contains a copy of a direct repeat (*a*) at the genome termini and also internally (*a'*, which is in an inverted orientation). Figure provided by A. Davison.

HHV-6A is highly neurotropic and it has been suggested it could be associated with neurological diseases such as multiple sclerosis (Berti *et al.,* 2000). HHV-7 has also been associated with exanthem subitum, and both HHV-6 and HHV-7 may act as opportunistic pathogens in immunocompromised individuals such as transplant recipients (Carrigan *et al.,* 1991).

Humans are the natural host for two members of the *Gammaherpesvirinae*: EBV and KSHV. Primary infection with EBV can result in infectious mononucleosis in adolescents, followed by clearance and persistent infection in B-lymphocytes. EBV has also been implicated in nasopharyngeal carcinoma, Burkitt's lymphoma and Hodgkin's lymphoma (Kutok and Wang, 2006). KSHV is the causative agent of Kaposi's sarcoma (KS), which is frequently found as a complication of HIV infection and in older men of Mediterranean or Eastern European background (Nascimento *et al.,* 2004). KSHV is common in Africa, where it infects patients of all ages but adult KSHV infects predominantly males. It causes more severe symptoms in children and young adults (Dedicoat and Newton, 2003, Dedicoat *et al.,* 2004; Wahman *et al.,* 1991).

1.3 Characteristics of HCMV

HCMV is a member of the genus *Cytomegalovirus* in the subfamily *Betaherpesvirinae*. Viruses in the genus are characterised by restricted host range, long life cycle in cell culture and the production of nuclear and cytoplasmic inclusions in infected cells. HCMV was first described in the 1930s in association with cytomegalic inclusion disease (CID) in infants, and the characteristic 'owl's eye' cytopathology was found in a number of cell types from infected patients. These enlarged cells (cytomegalia) resulted in the name cytomegalovirus.

HCMV was first identified as the causative agent of CID in the 1950s (Craig *et al.,* 1957; Rowe *et al.,* 1956; Smith, 1956).

1.3.1 Disease and epidemiology

HCMV can be transmitted via saliva, sexual contact, blood transfusion, transplantation (solid-organ and stem-cell), and also from mother to child via the placenta, during delivery or through breastfeeding. The virus is ubiquitous, infecting 50-90% of the worldwide population. Where studied in the developing world, the majority of people are infected at an early age, with seroprevalence approaching 90% (Gandhi and Khanna, 2004).

For the majority of those infected, primary HCMV infection is asymptomatic. Following initial lytic infection, the virus enters a latent state where it remains for the remainder of the host's life. However, HCMV infection can result in serious disease in a number of patient types. Congenital infection is one of the most important clinical manifestations of primary HCMV infection. Babies infected in the first trimester are most at risk, particularly if their mother is seronegative. Primary HCMV infections are reported in 1-4% of seronegative mothers during pregnancy, and transmission of the virus to the foetus occurs in 30-40% of these. Reactivation of the virus during pregnancy is reported in 10-30% of seropositive mothers and transmission of the virus to the foetus occurs in 1-3% of these. Of those infants infected congenitally, 5-10% develop irreversible symptoms, including hearing loss, encephalitis, visual impairment, mental retardation and sometimes death. HCMV can also be acquired perinatally and results in short-term, self-limiting symptoms in 30% of infants infected (Ahlfors *et al.*, 1982; Boppana e*t al.*, 1992; Malm and Engman, 2007; Stagno *et al.*, 1982, 1982a). Unfortunately, there is no treatment other than counselling; therefore an HCMV vaccine has been given priority by the US Institute of Medicine (Stratton, 2000).

HCMV infection is also a serious problem for immunocompromised individuals, including organ transplant recipients, allogeneic stem-cell recipients and AIDS patients. For transplant recipients, the combination of immunosuppression and post-operative stress can result in reactivation of HCMV that was latent in the recipient or the donor organ or cells.

Seronegative patients in receipt of a seropositive organ are most at risk, as they have no HCMV-specific immune response. It has also been suggested that increased viral load in the infected organ results in increased risk of HCMV disease. Infection initiates in the infected organ but rapidly spreads and can result in pneumonitis, enteritis, and hepatitis, and can potentially involve the CNS (Gandhi and Khanna, 2004). HCMV infection occurs in approximately 17% of stem-cell recipients and is usually due to reactivation of latent virus in a seropositive recipient (Zaia, 2002). Before the development of highly active antiretroviral therapy (HAART) for HIV-positive individuals, HCMV retinitis was found in 25% of patients. While this is no longer a problem in the developed world, there can be vitritis due to inflammation and encephalopathy due to replication of the virus in the CNS. In general, as AIDS progresses there is an increase in HCMV disease. It is not known whether this is a result of increasing immune dysfunction because of HIV progression or whether HCMV itself promotes progression of HIV (Gandhi and Khanna, 2004; Gerna *et al.,* 1998). In addition, HCMV infection in HIV positive children is associated with enhanced mortality (Kovacs *et al.,* 1999).

Diagnosis of HCMV infection was previously performed by cell culture of the virus. However, owing to the long replicative life cycle of HCMV, it can take weeks for visible plaques to form, depending on the inoculum (Drew, 1988). The availability of a monoclonal antibody (MAb) to IE protein p72 (IE1/UL123) allowed virus to be detected in infected fibroblasts within 24 h by fluorescence microscopy (Gleaves *et al.,* 1984). Diagnosis is now performed routinely by antigenaemia assay, ELISA, qualitative PCR and quantitative PCR. Diagnosis of congenital infection is performed by isolation of the virus from urine, or detection of virus DNA by PCR in urine, saliva, blood or cerebrospinal fluid. The presence of maternal immunoglobulin G (IgG) antibodies in the first three weeks of the child’s life indicates congenital infection. Diagnosis of prenatal congenital infection is more difficult and is usually by detection of HCMV DNA in the amniotic fluid (Malm and Engman, 2007; Revello and Gerna, 2004).

1.3.2 Immune response

The virion envelope contains a number of glycoproteins and among these, glycoprotein B (gB, UL55) is the predominant target for neutralising antibodies (Kari and Gerhz, 1990). Glyocproteins H (gH) and gO are also important targets for neutralising antibody, which prevents cell-to-cell spread (Paterson *et al.,* 2002, Urban *et al.,* 1996). The importance of the humoral response in HCMV infection is not fully understood. However it is thought that seronegative recipients of seropositive organs encounter more severe and more frequent primary infection because of the absence of HCMV-specific antibodies (Khanna and Diamond, 2006, Gandhi and Khanna, 2004). Natural killer (NK) cells are also thought to play a role in viral clearance through perforin-dependent cytolysis and non-cytolytically through induction of interferon (IFN) β (Iversen *et al.,* 2005). One patient with recurring severe HCMV disease was found to be deficient in NK cells (Biron *et al.,* 1989). It is notable that HCMV encodes at least six genes involved in NK evasion; UL16, UL142 and the microRNA, miR-UL112, prevent cell surface presentation of ligands for the NK cell activating receptor NKG2D, UL18 encodes a major histocompatibility complex class I (MHC-I) homologue, and UL40 upregulates expression of MHC-I E (also known as human leukocyte antigen E) (Tomasec *et al.,* 2000, 2005; Wilkinson *et al.,* 2008).

MHC-I-restricted HCMV-specific T-cell responses are important in the control of viral replication (Chen *et al.,* 2004). Previously it had been thought that CD8+ T cells against pp65 (UL83) or p72 (IE1/UL123) constitute the majority of T-cell responses in healthy carriers, but it is now known that they constitute only 40% (Day *et al.,* 2007). CD8+ T cells against other HCMV antigens, such as pp50 (UL44), gB (UL55), p86 (IE2/UL122), pp28 (UL99), pp150 (UL32), pp71 (UL82) and a number of proteins encoded in U_S, constitute the remaining 60% (Elkington *et al.,* 2003). Indeed, it is this broad repertoire of T-cell responses that establishes successful immune control of HCMV infection. A study in which donor-derived HCMV-specific CD8+ T cells were given to allogeneic stem-cells recipients resulted in immunity for the recipients, thus providing evidence for the importance of the T-cell response in HCMV immune defence (Walter *et al.,* 1995).

CD4+ T cells also play an important role in the control of HCMV infection. CD4+ T cells against pp65 (UL83), gB (UL55), gH (UL75), p72 (IE1/UL123), p86 (IE2/UL122) and UL69 have been found in some individuals. It is thought that in some patients the CD8+ T cell and antibody responses are insufficient to control primary infection and that effector-memory CD4+ T cells are required, perhaps to help maintain the virus-specific CD8+ T-cell response. Evidence for this was found when children with prolonged viral shedding showed persistent deficiency of the virus-specific CD4+ T-cell immune response and no deficiency in virus-specific CD8+ T cells (Chen *et al.,* 2004; Tu *et al.,* 2004).

HCMV has evolved a number of mechanisms for controlling the host immune response, including the down regulation of MHC-I and MHC-II molecules, the expression of MHC-I homologues, and NK evasion. HCMV also encodes homologues of interleukin 10 (IL-10), IL-8 and a number of chemokine receptors that could assist dissemination of the virus and prevent apoptosis. However, as discussed above, the virus does not avoid immune recognition and induces a broad range of CD8+ T cells. The finding that the recovery of HCMV-specific T-cell responses resulted in decreased chances of developing HCMV disease led to attempts to restore cellular immunity.

A number of methods to isolate HCMV-specific CD8+ T cells have been tried, including artificial antigen-presenting cells, peptide-pulsed dendritic cells and use of peptide-MHC-I tetramers. Experiments using adoptive transfer failed to elicit a virus-specific CD4+ T-helper response, but pre-emptive infusion of HCMV-specific CD8+ and CD4+ T cells resulted in expansion of the virus-specific T-cell response and reduced the incidence of HCMV disease in recipients (Walter *et al.,* 1995).

1.3.3 Treatment of infections

As there is no vaccine available, preventative measures such as hand washing are important and have resulted in reduced virus transmission in the developed world, particularly in child care centres, which were previously an area of increased risk of infection (Bale *et al.,* 1999). Antivirals such as ganciclovir, cidofovir and foscarnet (which inhibit the viral DNA polymerase (UL54)) are used to treat HCMV infection in immunocompromised individuals. Phosphorylated ganciclovir is a deoxyguanosine mimic that accumulates in infected cells and is incorporated into the growing DNA chain during viral replication; following its incorporation, one additional nucleotide is incorporated before the DNA polymerase stalls (Schaeffer *et al.,* 1978). Cidofovir mimics deoxycytidine monophosphate. It is dephosphorylated by cellular enzymes and incorporated into the growing DNA chain during viral replication, but it does not result in stalling of the viral DNA polymerase unless two cidofovir residues are incorporated sequentially (Xiong *et al.,* 1997). Foscarnet (phosphonoformic acid) is an analogue of pyrophosphate. Unlike ganciclovir and cidofovir it does not compete with deoxynucleoside triphosphates. Instead, it binds the site normally occupied by pyrophosphate and prevents normal pyrophosphate release, so that the polymerase cannot complete the catalytic cycle (Eriksson *et al.,* 1982).

Fomivirsen is an oligonucleotide that is complementary to MIE messenger RNA (mRNA) and inhibits translation (Azad et *al.,* 1993).

Owing to teratogenic effects, none of the antivirals described above are licensed for use in congenital infections (Faqi *et al.,* 1997). There have been a number of small-scale studies using ganciclovir in infants with congenital HCMV. However, although treatment suppressed viral replication temporarily, it did not prevent long-term damage (Whitley *et al.,* 1997). More recently, Kimberlin *et al.,* (2008) showed that treatment of symptomatic congenital HCMV with ganciclovir resulted in decreased clinical symptoms during treatment and improved outcome. However, viral load increased upon cessation of treatment.

Long term use of HCMV antivirals, particularly in AIDS patients, has resulted in the occurrence of antiviral resistance, and more recently there have been reports of antiviral resistance in transplant recipients (Lurain *et al.,* 2002). Resistance to ganciclovir is most commonly due to mutation in the viral phosphotransferase gene (UL97), which reduces phosphorylation of ganciclovir to the form required for inhibition of the viral DNA polymerase (Sullivan *et al.,* 1992). Mutations are also found in the DNA polymerase gene (UL54), and can confer resistance to both ganciclovir and cidofovir. Mutations within UL54 also confer resistance to forscarnet (Baldanti *et al.,* 2004). Single UL54 mutations can confer resistance to all three antivirals (Chou *et al.,* 2003). Resistance to other antivirals in development, such as benzimidazole ribonucleosides, have also been described (Krosky *et al.,* 1998). This propensity for mutations conferring resistance to antiviral treatment increases the urgency for the development of an HCMV vaccine.

There have been a number of attempts to develop a vaccine, the first using the highly passaged, attenuated Towne strain. This vaccine had few side effects and induced CD4+ and CD8+ T-cell immunity and antibody responses. It also reduced the severity of HCMV disease in renal transplant recipients. However, it did not prevent infection in seronegative women (Plotkin, 2001; Adler, 1995). The second was a recombinant vaccine combining parts of the genome from Towne and the low passage isolate Toledo. This is currently in clinical trials with seropositive individuals (Berstein *et al.,* 2002; Heineman *et al.,* 2006). However, the long-term safety of such vaccines is a major concern, particularly for pregnant women.

A potentially safer strategy is the use of subunit vaccines involving the most relevant antigens for vaccination. An attempt using gB (UL55) induced a strong neutralising antibody response but showed poor efficacy in preventing HCMV infection. A more recent proposal of combining gB with pp65 (UL83) may improve efficacy, as pp65 is a dominant cytotoxic T cell target (Pass *et al.,* 1999). Another proposal is the use of live viral vectors to deliver multiple HCMV antigens. Initial testing in seropositive individuals using a recombinant canarypox encoding gB (UL55) did not induce a strong gB-specific neutralising antibody response (Adler *et al.,* 1999). More recently, a canarypox encoding pp65 (UL83) induced strong CD4+ and CD8+ T cell responses, and another poxvirus vector encoding gB, pp65 and pp150 (UL32) is being tested (Berencsi *et al.,* 2001; Wang *et al.,* 2004).

1.3.4 Virion structure

HCMV has a typical herpesvirus structure, consisting of the dsDNA genome encased in an icosahedral capsid, surrounded by a proteinaceous tegument layer, and enveloped in a host cell-derived lipid bilayer, which contains numerous viral glycoproteins (Figure 1.2A). The capsid, which is 130 nm in diameter, consists of an icosahedral (T=16) lattice consisting of 161 capsomers, which are composed of two distinct units, 150 hexamers (hexons) and 11 pentamers (pentons) located at the vertices. The capsomers are linked by triplex complexes (Figure 1.2B). The hexamers are 15.8 nm apart (centre-to-centre) at the outer edge in the HCMV capsid. The average diameter of the HCMV scaffold is 76 nm.

By analogy with HSV-1, the major capsid protein (MCP; UL86) forms hexons and pentons and the smallest capsid protein (SCP; UL48A) is located at the tips of the hexons. The minor capsid protein (mCP; UL85) associates with the mCP-binding protein (mC-BP; UL46) in a 2:1 ratio to form the triplex structures that link the capsomers (Bhella *et al.,* 2000; Butcher *et al.,* 1998). UL104 encodes the portal protein, which forms a high-molecular weight complex, the portal complex (Dittmer *et al.,* 2005). By analogy with HSV-1, the portal complex is located at one of the vertices and the HCMV genome is translocated into a pre-formed procapsid through it (Chang *et al.,* 2007).

Several capsid forms have been identified in the nuclei of infected cells by electron microscopy and sucrose gradient centrifugation. Type A capsids are devoid of DNA and are thought to be the result of abortive packaging. Type B capsids also lack DNA but contain viral scaffolding protein and are located in the nucleus (Gibson, 1996). Type C capsids contain the viral genome and can mature into infectious virions (Homa and Brown, 1997).

The capsid is assembled initially through the formation of an internal protein scaffold. The scaffold is composed of the viral protease (encoded by UL80) and the assembly protein (AP, encoded by UL80.5) (Wood *et al.,* 1997; Varnum *et al.,* 2004). The AP forms a scaffold by self-interaction via N-terminal sequences and interacts with MCP via C-terminal sequences (Oien *et* al., 1997). Proteolytic processing of these proteins, removal of the scaffold and packaging of viral genome in the core are essential for capsid maturation and the production of infectious virions.

In contrast to the capsid proteins, all of which have homologues in other mammalian herpesviruses, some tegument and envelope proteins have homologues only in other betaherpesviruses or a subset thereof. The tegument appears to be a largely amorphous proteinaceous coating of the capsid that maintains association between the capsid and the envelope. It has been suggested that, due to interaction with the capsid, the innermost tegument layer exhibits icosahedral symmetry (Chen *et al.,* 1999). The tegument contains at least 27 proteins, and the majority are phosphophorylated and highly immunogenic (Britt and Boppana, 2004). Tegument proteins perform a diverse range of functions from transcriptional activation (UL26, Stamminger *et al.,* 2002) to cell cycle progression (UL82, pp71, Kalejta and Shenk, 2003) to envelopment (UL99, pp28, Silva *et al.,* 2003). It may be that the majority of tegument proteins function specifically within the infected cell. In this respect, the tegument can be viewed both as a part of the virion structure and as a system for delivering viral proteins to the cell immediately upon infection.

The envelope contains eight experimentally confirmed glycoproteins, but as many as 40 genes potentially encode glycoproteins, some of which may be

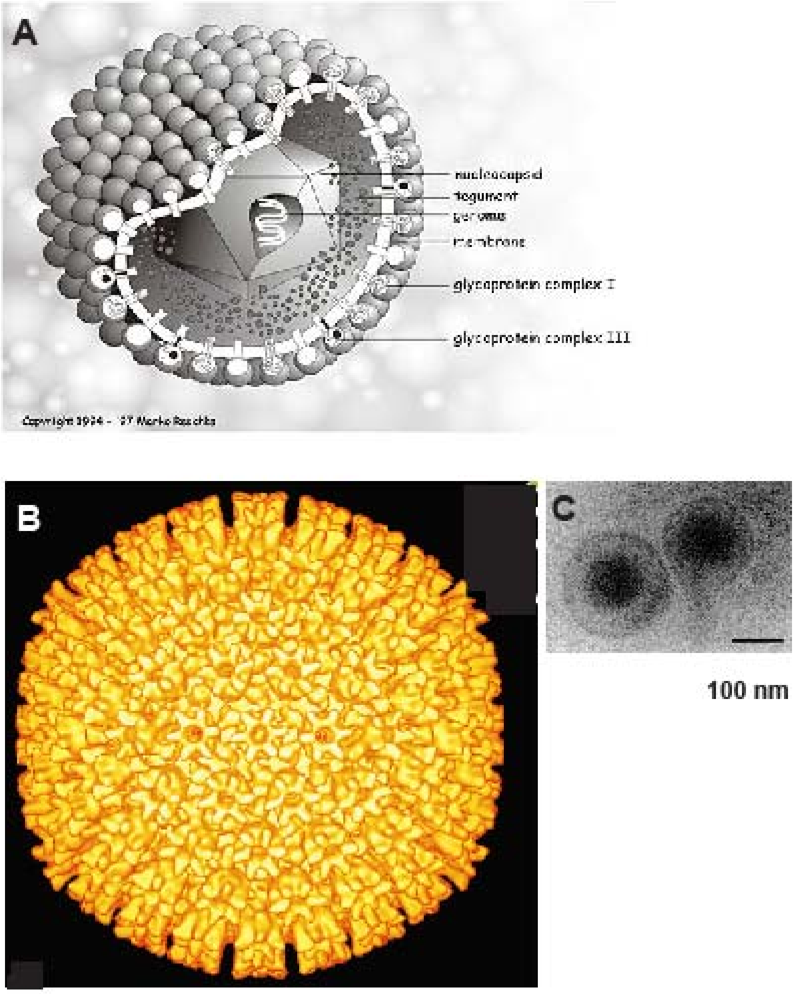


Figure 1.2 HCMV virion and capsid structure

A) Cartoon representation of an HCMV virion. The genome is encased in an icosahedral capsid, surrounded by tegument, within a lipid bilayer envelope, which contains viral glycoproteins. From Reschke, http://www.virology.net/Big_Virology/BVDNAherpes.html.

B) Cryo-electron micrograph reconstruction of the surface of an HCMV capsid (from Butcher *et al.,* 1998).

C) Cryo-electron micrograph of HCMV virions. From Bhella *et al.,* (2000).

present in the envelope (Chee *et al.,* 1990). Some of the more abundant glycoproteins such as gB (UL55), gM/gN (UL100/UL73) and gH/gL/gO (UL75/UL115/UL74) exist as disulfide-linked complexes (gCI, gCII and gCIII, respectively) within the virion Figure 1.2A. All of these have been shown to be essential for production of infectious virus. A recent mass spectrometry-based analysis of the relative abundance of HCMV virion proteins confirmed that the most abundant virion protein is pp65 and showed that the predominant glycoprotein is gM. Virion preparations were found to contain 71 host cellular proteins, including cytoskeletal proteins, proteins involved in transcription initiation and elongation, structural proteins, enzymes and chaperones (Baldick and Shenk, 1996; Britt and Boppana, 2004; Gretch *et al.,* 1988; Hobom *et al.,* 2000; Varnum *et al.,* 2004).

1.3.5 Genome structure

HCMV has the largest known human herpesvirus genome. Its linear dsDNA genome is approximately 236 kbp in length and is predicted to contain approximately 165 genes (Dolan *et al.,* 2004). The genome consists of U_L and a unique short region (U_S) both flanked by inverted terminal repeats (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 (Figure 1.1). The genome also possesses a short region (called the *a* sequence) present as a direct repeat at its termini and also in inverse orientation at the IR_L-IR_S junction (Spaete and Mocarski, 1985). U_L and U_S can invert relative to each other resulting in four different isomers in virions (Mocarski and Courcelle, 2001).

1.3.6 HCMV genetic content

The first HCMV genome to be sequenced was the highly passaged, commonly used laboratory strain AD169 (229,354 bp) (Chee *et al.,* 1990). Subsequently, sequence errors were found: two in UL102, one of which results in an extension of the 5'-end (Smith and Pari, 1995), and one in US28 that results in extension of the 3'-end (Neote *et al.,* 1993). Sequencing of the right end of U_L in Towne and the low passage strain Toledo revealed that AD169 is a multiple mutant.

It contains a 15 kbp deletion at the right end of U$_L$, which contains 19 additional ORFs in Toledo. The deleted sequences have been replaced by an inverted sequence from the left end of the genome, resulting in an expansion of R$_L$ in AD169 (RL1-RL12 and part of RL13) (Cha *et al.,* 1996). AD169 also contains frameshift mutations in genes RL5A, RL13 and UL131A (Akter *et al.,* 2003; Davison *et al.,* 2003, 2003a; Yu *et al.,* 2002). Some stocks of AD169 were found to contain additional mutations in UL42 and UL43 or UL36 (Dargan *et al.,* 1997; Mocarski *et al.,* 1997; Skaletskaya *et al.,* 2001). In summary, AD169 differs significantly from wild-type HCMV strains. Moreover, the low passage strain Toledo retains sequences at the right end of U$_L$ but a substantial region is inverted in comparison to clinical HCMV isolates (Davison *et al.,* 2003; Lurain *et al.,* 1999). More recently, the low passage clinical isolate Merlin has been sequenced and is thought to represent wild type HCMV apart from a point mutation in UL128 that results in premature termination (Figure 1.3, Dolan *et al.,* 2004). This suggests that even a small number of passages in human fibroblasts can result in mutation of the virus genome.

A major difference between clinical HCMV isolates and highly passaged laboratory strains is their ability to infect different cell types in culture. Clinical isolates can infect fibroblasts, epithelial and endothelial cells whereas laboratory strains (invariably passaged in fibroblast cells) lose the ability to infect or replicate in epithelial and endothelial cells.

This change in tropism is associated with mutation in one of three genes in the UL128 locus (UL128, UL130 and UL131A) (Akter *et al.,* 2003; Gerna *et al.,* 2005; Hahn *et al.,* 2004). Although the UL128 locus is detrimental to growth in fibroblast cells, it is required for growth in epithelial and endothelial cells (Hahn *et al.,* 2004). Mutations have also been described in genes RL5A, RL13 and UL9 (all related members of the RL11 family) following fibroblast adaptation of clinical isolates, and suggests roles in tropism for these genes (Dolan *et al.,* 2004). The function of these proteins is not known, but another member of the RL11 family (RL11) has been shown to bind the Fc domain of IgG (Atalay *et al.,* 2002). Mutational analysis of a bacterial artificial chromosome (BAC) of Towne has suggested roles for UL9 in cell tropism (Dunn *et al.,* 2003).

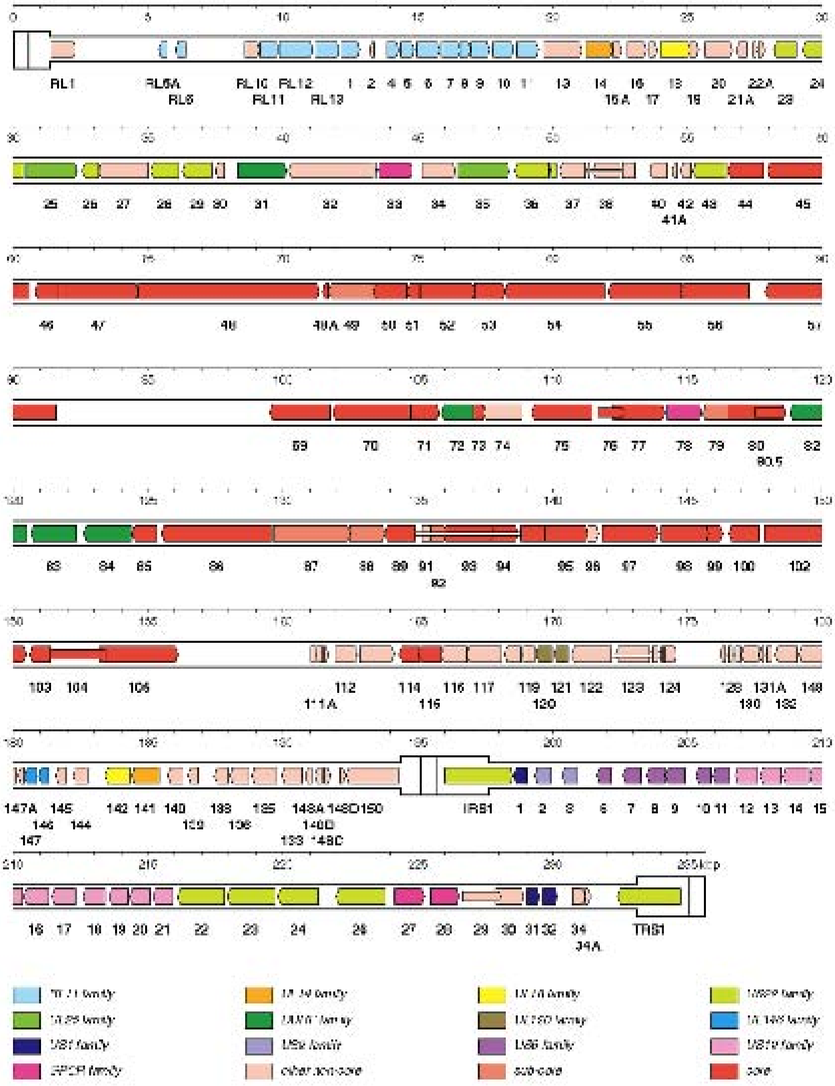

RL1
RL5A
RL6
RL10
RL11
RL12
RL13
1 2 4 5 6 7 8 9 10 11 13 14 15A 16 17 18 19 20 21A 22A 23 24
25 26 27 28 29 30 31 32 33 34 35 36 37 38 40 41A 42 43 44 45
46 47 48 48A 49 50 51 52 53 54 55 56 57
69 70 71 72 73 74 75 76 77 78 79 80 80.5 82
83 84 85 86 87 88 89 91 92 93 94 95 96 97 98 99 100 102
103 104 105 111A 112 114 115 116 117 119 120 121 122 123 124 128 130 131A 132 148
147A 145 146 144 147 142 141 140 139 138 136 135 148A 148D 150 133 148B 148C
IRS1 1 2 3 6 7 8 9 10 11 12 13 14 15
16 17 18 19 20 21 22 23 24 26 27 28 29 30 31 32 34 34A TRS1
235 kbp
US1 family
US2 family
US6 family
GPCR family
other non-core
sub-core
core

Figure 1.3 Genome map of HCMV strain Merlin

U_L and U_S are bounded by terminal and internal repeats (thicker regions), which contain the *a* sequence as a direct repeat at the genome termini and as an inverted repeat at the junction of IR_L and IR_S. Introns are shown as narrow white bars and protein-coding regions are shown as coloured arrows with the ORF name below. The 'core' genes common to alpha-, beta- and gammaherpesviruses are shown in red, whereas the subcore genes (found in beta- and gammaherpesviruses only) are shown in pink. Noncore genes are grouped into gene families and are coloured according to their gene families. Figure provided by A. Davison.

1.4 Replication cycle

A simplified diagram of the HCMV life cycle is shown in Figure 1.4. By analogy with other herpesviruses, the three major HCMV glycoprotein complexes, gCI, gCII and gCIII are thought to mediate attachment and entry via initial binding to a heparan sulfate cell receptor. gCI (gB) is the major heparan sulfate-binding protein and is thought to be necessary for entry into all cell types (Kari and Gehrz 1992). Additional cell surface components have been identified as HCMV receptors, including epidermal growth factor receptor, integrin $\alpha v\beta 3$ (Wang *et al.,* 2003), platelet-derived growth factor-α receptor (Soroceanu *et al.,* 2008) and toll-like receptor 2 (Compton, 2004). Binding initiates a cascade of events that results in fusion of the viral envelope with the cell membrane and release of the capsid and tegument into the cytoplasm. The other glycoprotein complexes may facilitate entry into various cell types. The gCII complex (gN/gM:UL73/UL100) also facilitates entry into the cell by binding heparan-sulfate. The gCIII complex (gH/gL/gO: UL75/UL115/UL74) facilitates entry into fibroblasts by fusion, whereas gH/gL complexed with proteins encoded by the UL128 locus may promote entry into epithelial and endothelial cells by endocytosis in a pH-dependent manner (Paterson *et al.,* 2002, Ryckman *et al.,* 2008; Singzer *et al.,* 2008). Alternative entry pathways have also been described for other herpesviruses, such as HSV-1 and EBV (Hutt-Fletcher, 2007; Nicola *et al.,* 2005). After penetration, the HCMV capsid is transported along microtubules to the nucleus, which the viral genome enters through a nuclear pore (Dohner *et al.,* 2005, 2005a). The tegument protein UL48 and the binding protein UL47 are both essential for replication (Dunn *et al.,* 2003) and are thought to control uncoating and release of viral DNA at the nuclear pore (Bechtel and Shenk, 2002).

Once the genome enters the nucleus, viral genes are expressed in a temporal cascade, the first genes expressed being termed the immediate-early (IE) or α genes, followed by the early (E) or β genes, and then finally by the late (L) or γ genes. IE genes are expressed immediately upon cell entry and do not require expression of other viral genes. E genes require IE protein products for expression and can be subdivided into E or β_1 and delayed-early (D-E) or β_2, which are expressed at slightly differing times. L genes can also be subdivided into two subclasses, early-late (E-L) or γ_1 and true L or γ_2, which differ in their dependence upon viral DNA replication (Mocarski and Courcelle, 2001). Indeed, only a few viral genes are true L, where expression depends absolutely on viral DNA synthesis (Mocarski and Courcelle, 2001; Spector, 1996). E genes tend to encode proteins involved in viral DNA replication or modulation of the host cell and host immune response. L genes tend to encode structural proteins.

The tegument protein pp71 (UL82) is thought to play an important role in the control of IE gene expression. The mechanism by which it does this is unclear, although the interaction between pp71 and the cellular protein hDaxx is thought to be involved. It has been suggested that hDaxx allows translocation of pp71 to nuclear domain 10 (ND10), a site of viral IE transcription (Ishov *et al.*, 2002). Other studies have suggested that pp71 relieves hDaxx-mediated repression of major immediate early (MIE) expression (Cantrell and Bresnahan, 2006). The pp71 protein is also involved in cell cycle control. It accelerates transition from G_0 to G_1 and progression through G_1, the latter by binding members of the retinoblastoma (Rb) family and promoting their degradation by a novel proteasome-dependent, ubiquitin-independent mechanism (Kalejta *et al.*, 2003). The most abundantly expressed IE genes are transcribed from the MIE locus and are IE1/UL123 and IE2/UL122 (Stenberg, 1996). Transcription occurs from a single transcription start site, and differential splicing and polyadenylation result in mRNAs whose products play an important role in the regulation of viral and cellular gene expression. Following expression of the MIE genes, the rest of the genome becomes transcriptionally active.

Viral DNA replication results in the formation of head-to-tail concatameric genomes that need to be cleaved into unit-length genomes for packaging. First, the DNA undergoes site-specific cleavage at *pac* motifs within the *a* sequence (Spaete and Mocarski, 1985). Unit-length genomes are then encapsidated into preassembled capsids. This process is catalysed by a virus-encoded enzyme complex called the terminase, in an ATP-dependent manner. The terminase is composed of two proteins, pUL56 and pUL89 (Bogner *et al.*, 1998; Giesen *et al.*, 2000; Spaete and Mocarski, 1985). Encapsidation of the HCMV genome is achieved via a larger capsid volume and a higher packaged DNA density than those of other herpesviruses (Bhella *et al.*, 2000, Butcher *et al.*, 1998).

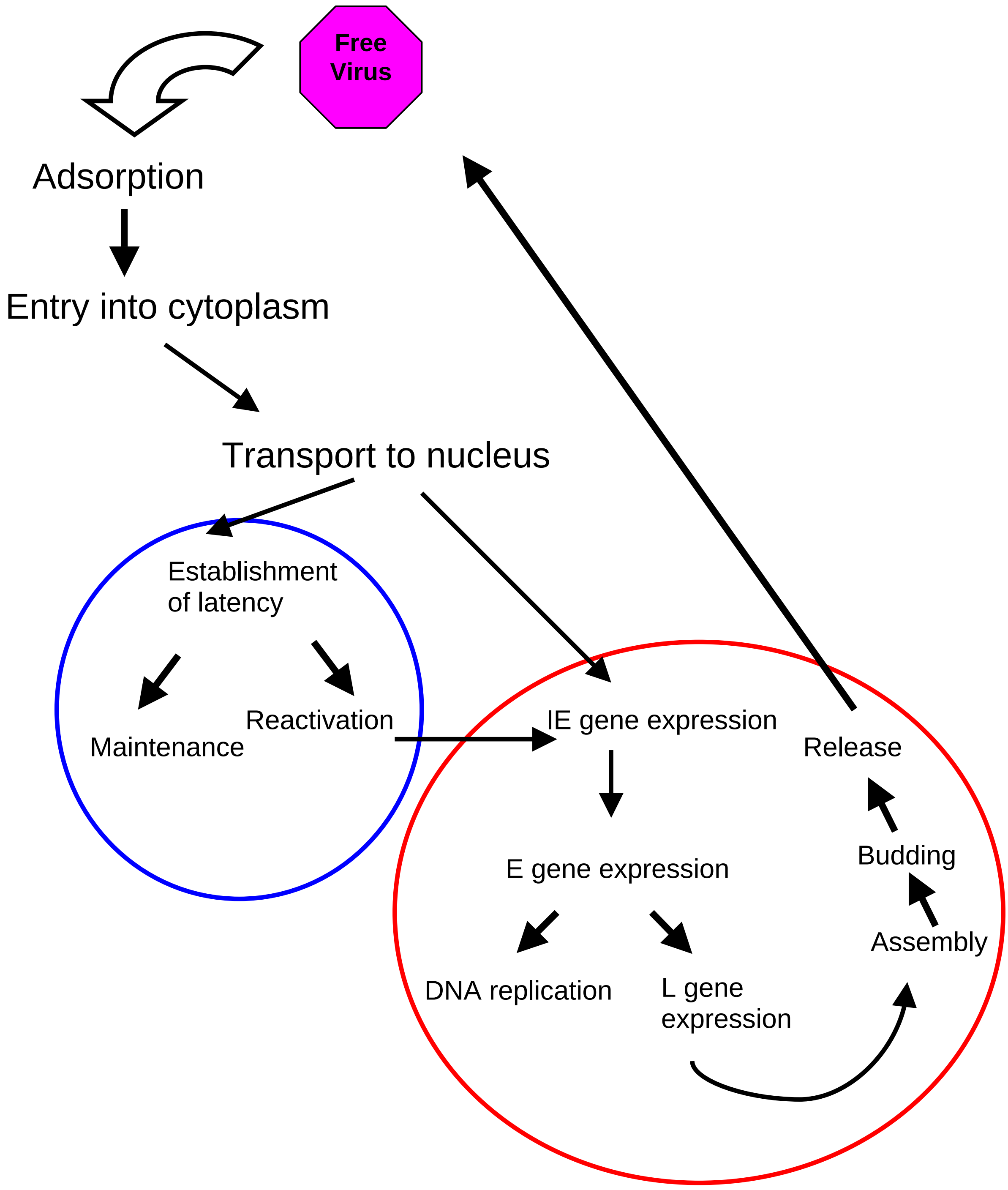


Figure 1.4 Replicative life cycle of HCMV
Once the virus enters the cell it is transported along microtubules into the nucleus, where it can establish latency (blue circle) or undergo productive infection (red circle).

The process of herpesvirus envelopment and egress from the nucleus is complex. The most commonly favoured model involves envelopment at the inner nuclear membrane followed by de-envelopment at the outer nuclear membrane, resulting in release of the capsid into the cytoplasm (Homman-Loudiyi *et al.,* 2003; Muranyi *et al.,* 2002; Mettenleiter, 2004). The tegument layer is then added in the cytoplasm via a complex system of protein-protein interactions. This is followed by secondary envelopment, which occurs by budding of the tegumented capsid into vesicles of the trans-Golgi network (Homman-Loudiyi *et al.,* 2003) or post-trans-Golgi endocytic membranes (Fraile-Ramos *et al.,* 2002). Mature virions are then transported to the cell surface using the cellular exocytic pathway. Both tegumentation and envelopment are mediated by specific protein-protein interactions (Mettenleiter, 2004).

1.5 Latency

Like other herpesviruses, HCMV establishes lifelong latent infection following primary lytic infection. The virus establishes latency at specific sites in the host without production of infectious virus and hence avoids immune recognition. For many years the exact site of HCMV persistence eluded detection. It was not until the advent of PCR that this question was addressed. HCMV DNA was detected by nested PCR combined with fluorescent-activated cell sorting (FACS) in the peripheral blood monocytes (specifically CD3- or non-T cells) of healthy, seropositive individuals (Taylor-Wiedemen *et al.,* 1991).

HCMV DNA has also been detected in CD34+ myeloid progenitor cells in the bone marrow (Mendelson *et al.,* 1996). Interestingly, CD34+ cells give rise to monocytes, as well as to other cell types, including B cells, T cells and polymorphonuclear leukocytes (PMNLs). To date, HCMV DNA has not been detected in B cells, T cells or PMNLs. However, it has also been detected in CD14+ monocytes, dendritic cells (DCs) and megakaryocytes.

The viral genome isolated from peripheral blood monocytes migrates as a circular plasmid on native agarose gels, suggesting that it is maintained as a circular episome (Bolovan-Fritts *et al.,* 1999). No evidence for viral IE expression has been found, consistent with a situation in which HCMV can be carried in a true latent state (Taylor-Wiedemen *et al.,* 1994).

Monocytes are non-permissive for viral replication, and it is only when they undergo differentiation into differentiated macrophages and immature DCs that productive infection is permitted (Sinclair and Sissons, 2006). Reactivation of viral gene expression, albeit in the absence of production of infectious virus, has been demonstrated by *in vitro* differentiation of monocytes (Taylor-Wiedemen *et al.,* 1994). Supplementation with medium containing cytokines resulted in complete reactivation of infectious virus (Söderberg-Nauclér *et al.,* 2001). More recently, *ex vivo* differentiation of CD34+ myeloid progenitors to mature DCs resulted in complete reactivation of infectious virus (Reeves *et al.,* 2005).

It is not known how the latent genome is maintained or whether HCMV encodes latent genome maintenance factors corresponding in function to EBV nuclear antigen 1 (EBNA-1). However, deletion of sequences near the MIE locus affected maintenance of HCMV genomes in experimentally infected undifferentiated granulocyte-macrophage precursors (GMPs) maintained in long-term cell culture (Mocarski, 2006). One interesting theory is that there is no viral replication in CD34+ myeloid progenitors, rather the HCMV genome is carried passively by these cells until they differentiate into macrophages and DCs where it then reactivates and is reseeded into peripheral blood monocytes (Sinclair and Sissons, 2006).

An experimental model system of latency, using experimentally infected GMPs derived from foetal liver cells, was developed by Kondo *et al.* (1994). Using this system, several HCMV transcripts expressed in the absence of productive infection were identified. These transcripts, termed CMV latency-specific transcripts (CLTs), included novel spliced and unspliced RNA transcripts that map to the MIE locus. They have been detected in healthy seropositive individuals as well as during cell culture suggesting a role in latency (Kondo *et al.,* 1994; Kondo and Mocarski, 1995).

Jenkins *et al.* (2004) detected transcription from the UL111A locus using the same experimental model of latency. UL111A encodes a protein that is homologous to the immune modulator IL-10, termed viral IL-10 (vIL-10), which is expressed during productive infection (Kotenko *et al.,* 2000). The UL111.5A transcript detected during latency displays an alternative splicing pattern, which results in premature termination of the protein. Jenkins *et al.* (2004) also detected this incompletely spliced transcript in monocytes and peripheral blood cells of healthy carriers, suggesting that it is expressed during natural latent infection.

Another study detected an antisense RNA in the bone-marrow monocytes of seropositive individuals that is antisense to the UL82 gene, which encodes pp71, a known transactivator of the MIE locus (Bego *et al.,* 2005). This could indicate a role for latent transcripts in restriction of gene expression. Yet another study suggested a role for histone deacetylase in repression of the MIE locus in non-permissive cells, and found that the MIE promoter associates with heterochromatin protein 1 (HP1) in peripheral blood monocytes. HP1 is a chromosomal protein that has been implicated in gene silencing (Murphy *et al.,* 2002). Further studies are required to elucidate the true mechanism of HCMV latency and subsequent reactivation, particularly in immunocompromised individuals where it has serious consequences.

1.6 Variability in HCMV

Whole genome comparisons of sequences from different HCMV strains have shown that the genome is highly conserved between strains at both the nucleotide and imputed amino acid (aa) sequence levels (>95%) (Murphy *et al.,* 2003; Dolan *et al.,* 2004). However, highly polymorphic regions are dispersed throughout the genome in both coding and noncoding regions, with aa sequence identity of 50-80% in the former. Many hypervariable genes are predicted to encode glycoproteins that are potentially expressed on the surface of infected cells and possibly also embedded in the virion envelope, thus making them potential targets for the immune system.

Figure 1.5 shows the nucleotide divergence between nine HCMV strains, which was calculated using an alignment of the sequences at the right end of U_L. UL146 and UL139 are the two most variable genes in this region (Dolan *et al.,* 2004). UL146 and UL139 are the focus of this thesis and will be discussed in more detail in later sections (Sections 1.9 and 1.10).

The following section provides a brief description of gene variation in HCMV, focussing specifically on hypervariable genes, potential linkage disequilibrium strains

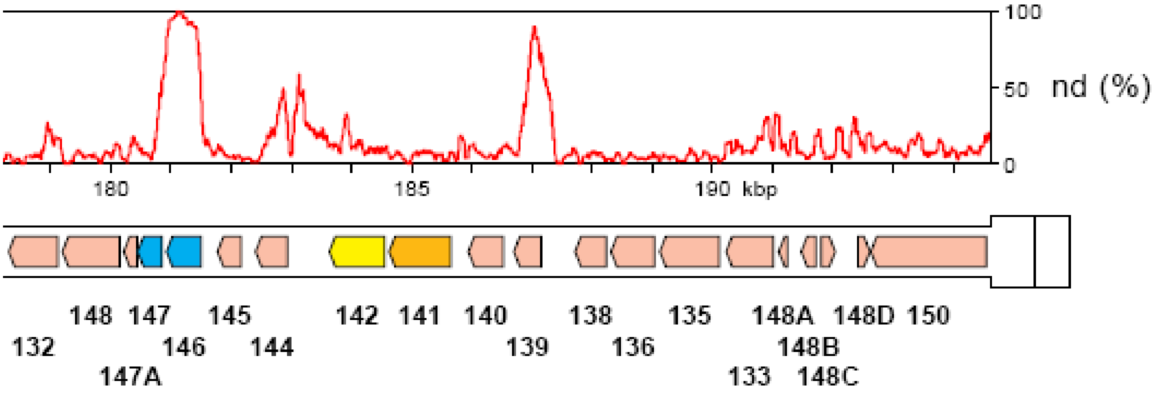


Figure 1.5 Nucleotide divergence at the right end of U_L in nine HCMV strains
Nucleotide divergence between nine strains (Davis, Towne, Toledo, TB40/E, Merlin, 3157, 6397, 3301 and W) was calculated using an alignment of the sequences at the right end of U_L. If all strains were not identical, a nucleotide position was counted as divergent, as were gaps in the alignment. The inversion in Toledo was corrected. The plot shows nucleotide divergence in a 100 nucleotide window shifted by increments of three nucleotides. The protein-coding regions in this region of the genome are shown below the plot, with a scale based on their position in strain Merlin. From Dolan *et al.* (2004).

between hypervariable loci, and association between genotype and disease outcome. Linkage disequilibrium is a term that describes the non-random association of alleles, or genotypes, at two or more loci. The information is summarised in Table 1.2. In addition, the occurrence of mixed HCMV infections, where more than one genotype is detected in the same sample, is described below, and also in more detail in Section 1.7.

Currently, there is no universally accepted definition of what constitutes a genotype. For the purpose of this study, a genotype is defined by phylogenetic analysis with bootstrapping, where all sequences in a genotype cluster tightly together and nucleotide and amino acid identity is high (>97%). This differs from some groups' interpretation, particularly He *et al.,* 2006, who describe five groups for UL146 rather than the 14 genotypes described by Dolan *et al.,* 2004 and the present study. He and colleagues grouped strains from different genotypes together, resulting in aa and nt identities below 80%.

1.6.1 RL11 family

The RL11 family contains 14 members: RL5A, RL6, RL11-RL13, UL1, and UL4-UL11. All members are in close proximity on the HCMV genome, and all but UL5 and UL8 contain the characteristic RL11D domain. This domain consists of a region of variable length (65-82 aa) containing three conserved aas (W, C and C) and a number of potential N-linked glycosylation sites. Most RL11 proteins are believed to encode transmembrane glycoproteins, although their functions have yet to be determined. Several RL11 family members are hypervariable, particularly RL12, RL13 and UL9 (Dolan *et al.,* 2004). Proteins containing the RL11D domain have also been identified in members of the family *Adenoviridae*, in the E3 region (Davison *et al.,* 2003, 2003a; Dolan *et al.,* 2004).

Sekulin *et al.* (2007) analysed the sequences of the RL11D domain in several RL11 family genes (UL1, UL4, UL6, UL7 and UL10) in 70 unpassaged clinical isolates, and confirmed them as highly variable. UL1, UL7 and UL10 fell into three genotypes whereas UL4 and UL6 fell into four genotypes. UL1 showed the highest level of variation and in addition ~13% of samples contained frameshifts or point mutations that resulted in a stop codon and premature truncation of UL1. Multiple genotypes were detected in 28 samples (40%). Statistically significant linkages between the genotypes of the genes examined were detected, which may be a consequence of their close proximity on the genome. Specifically, evidence for linkage disequilibrium was found between UL6 and UL7, UL4 and UL7, UL1 and UL4, and UL4 and UL6. As the clinical samples were obtained from a number of body sites, investigation of potential compartmentalisation was performed. No significant association was found with the exception of one UL7 genotype (B), which was detected exclusively in urine samples at the 5% significance level (Sekulin *et al.,* 2007).

1.6.2 UL4 major transcript leader

UL4 is a member of the RL11 family and encodes a glycoprotein (gpUL4 or gp48) (Chang *et al.,* 1989; Dolan *et al.,* 2004). Expression of UL4 is controlled by an unusual translational mechanism, where the peptide product of a small ORF (uORF2) upstream of UL4 (within the 5'-leader sequence) apparently blocks translation termination at its own stop codon and causes ribosome stalling. This prevents other ribosomes from accessing the UL4 initiation codon. uORF2 has been shown to be hypervariable in the N-terminal region and this sequence variation results in variation in repressor activity (Alderete et *al.,* 1999).

A study by Bar *et al.* (2001) investigating UL4 leader sequences in ten AIDS patients and 21 bone marrow transplant recipients described four genotypes (1, 2, 3A and 3B) with 23% nucleotide divergence between sequences. Polymorphisms were dispersed throughout the leader sequence. More than one strain was found in five of the ten patients. In all but one patient with a mixed infection, different genotypes were isolated from different body sites, but genotype 3B, found in a single patient, was isolated from four different body sites, suggesting no association between genotype and tissue sample. The same study found no evidence for linkage disequilibrium between UL4 genotypes and gB genotypes (Bar *et al.,* 2001).

1.6.3 UL11

UL11 is a member of the RL11 family (Chee *et al.,* 1990). A study by Hitomi *et al.* (1997) investigated UL11 sequences in eight passaged clinical strains and compared them with UL11 in Towne and AD169. UL11 was found to be highly variable towards the N-terminus, which is predicted to contain a signal sequence and an extracellular domain (Chee *et al.,* 1990), whereas the C-terminal region was highly conserved. All sequences fell into three genotypes, with aa sequence identity only 57% in the N-terminal region (Hitomi *et al.,* 1997).

1.6.4 UL37

The UL37 locus encodes three UL37 IE proteins, the UL37 exon 1 protein (pUL37x1), pUL37 (which is encoded mostly by UL37 exon 3) and pUL37$_M$. All three proteins contain the same N-terminal signal sequence, a strongly charged acidic domain and two domains essential for their anti-apoptotic activity (encoded by UL37 exon 1). pUL37 and pUL37$_M$ are both N-linked glycoproteins, which are produced via alternative splicing and polyadenylation of UL37 RNA (Goldmacher *et al.,* 1999). Investigation of UL37 exon 1 sequences in 26 HCMV strains, four of which were passaged, found them to be highly conserved (Hayajneh *et al.,* 2001a). UL37 exon 3 encodes the C terminus of pUL37 and pUL37$_M$, and was found to be variable in 20 clinical isolates (15 unpassaged and five passaged once on HFFs), with variation concentrated in the first three-quarters of the exon (aa sequence divergence of 28%). In contrast, residues within the transmembrane region and cytosolic tail were highly conserved (Hayajneh *et al.,* 2001).

1.6.5 UL55

UL55 encodes gB, which is essential for viral replication *in vivo* and *in vitro*, and has roles in virus attachment, cell entry and cell-to-cell spread (Section 1.4). Expression of UL55 results in a glycosylated precursor molecule that is cleaved after codon 461 to generate gp55 and gp116, which together form a dimeric complex called gCI through the formation of disulfide bonds (Britt and Vulger, 1989). Variation in gB was first described by Chou and Dennison (1991), when restriction fragment length polymorphism (RFLP) analysis and partial sequencing revealed that HCMV strains fell into four main genotypes (gB1-gB4), based on variation around the cleavage site of gB. Variation towards the N terminus has also been described, whereas the C terminus is well conserved. Investigation of sequences at these three sites (i.e. the cleavage site, the N terminus and the C terminus) confirmed four genotypes (termed gBn1/gBcls1/gBc$_{1/2}$, gBn2/gBcls2/gBc$_{1/2}$, gBn3/gBcls3/gBc$_{3/4}$ and gBn4/gBcls4/gBc$_{3/4}$) (Meyer-König *et al.*, 1998).

Three additional genotypes (gB5-gB7) have been detected since, albeit at very low frequencies, and these may actually be subtypes of gB1 and gB3 (Shepp *et al.*, 1998; Trincado *et al.*, 2000). The full degree of sequence divergence between gB genotypes has not been described, as most gB genotyping studies relied on RFLP and partial sequences. Some evidence has been presented for differences in geographical distribution of gB genotypes (Zipeto *et al.*, 1998), and mixed infections have been detected (Aquino and Figueirdo, 2000). Investigation into potential differences in cell tropism between gB genotypes found that strains with gB1 did not infect T lymphocytes whereas those with gB2 and gB3 did. However, this may reflect small sample size (ten), as all gB genotypes were detected in blood and urine samples (Meyer-König *et al.*, 1998), which is in agreement with the findings of Carraro and Granato (2003).

There have been numerous studies that utilised gB genotyping to investigate potential correlation with disease. In one study, gB1 was more commonly detected in bone marrow transplant recipients with non-fatal HCMV infection compared with fatal cases (Fries *et al.*, 1994). AIDS patients who developed retinitis as a complication of HCMV infection were more frequently infected with gB1 than other genotypes (Rasmussen *et al.*, 1997). Analysis of gB sequences from 15 congenital infections in Hungarian samples found that they all contained gB1 (Lukacsi *et al.*, 2001). In contrast, investigation of gB in renal transplant recipients (n=34) revealed no association between gB genotype and the development of HCMV disease (Aquino and Figueirdo, 2000). Contradictory findings may reflect the fact that most of these studies used small sample sizes and different patient types. This, together with lack of linkage between gB genotypes and the genotypes of other variable genes, means that any conclusions need to be treated with caution.

1.6.6 UL73

UL73 encodes a type I transmembrane glycoprotein, gN, which together with gM/UL100 forms the glycoprotein complex gCII, which is a major heparin-binding complex (Section 1.4).

gN is a major target for the immune system and induces a neutralising antibody response. gM is thought to act as a chaperone for gN processing, and is highly conserved (Pignatelli *et al.,* 2004). In contrast, gN is highly variable, with variation concentrated in the highly glycosylated N-terminal region. Phylogenetic analysis of 40 UL73 sequences from clinical isolates revealed four major genotypes (gN1-gN4), with gN4 divided into three subtypes (gN4a, gN4b and gN4c) and nucleotide sequence identity ranging from 80-87%. UL73 sequences were found to be stable over time within patients and when passaged in cell culture (Pignatelli *et al.,* 2001).

A large scale study by Pignatelli *et al.* (2003), which examined UL73 sequences in 223 clinical samples (urine and saliva samples were passaged, whereas all other samples were unpassaged) from a number of locations worldwide, confirmed the existence of four main genotypes and also identified a novel subgroup within gN3, which resulted in division of gN3 into two subtypes (gN3a and gN3b). This study also investigated potential bias in the geographical distribution of genotypes and the possibility that these sequences were under positive selection. The genotypes gN1 and gN2 were found to have evolved under neutral selection, whereas gN3 and gN4 each contain regions that are under positive selection. No differences in genotypic frequencies between regions were observed and all gN genotypes were represented in all regions. Perhaps as a consequence of a much larger sample size, aa sequence divergence was found to be as high as 50% between some sequences.

Dal Monte *et al.* (2004) considered whether there was any linkage between gN genotype and cellular tropism. They hypothesised that isolates collected from urine and saliva samples had epithelial cell tropism and isolates collected from blood or biopsy samples had endothelial cell tropism. They analysed UL73 sequences in 102 samples with endothelial cell tropism and in 81 samples with epithelial cell tropism. However, no significant association between gN genotypes and epithelial or endothelial tropism was found.

1.6.7 UL74

The envelope glycoprotein complex gCIII (which is composed of gH/gL/gO) mediates cell entry and cell-to-cell spread through membrane fusion (see

Section 1.4). gH (UL75) and gL (UL115) are relatively well conserved between HCMV strains, although two gH genotypes (gH1 and gH2) have been described (Chou, 1992; Pignatelli *et al.,* 2004). In contrast, gO (UL74) is hypervariable, particularly towards the N terminus where divergence reaches 40% (Paterson *et al.,* 2002).

Both gH (UL75) and gL (UL115) are found in all mammalian and avian herpesviruses studied to date, whereas gO is specific to betaherpesviruses. EBV also encodes a gH/gL complex that includes a third component, in this case gp42, which is unrelated to gO. gp42 is essential for infection of B lymphocytes but is not required for infection of epithelial cells (Wang and Hutt-Fletcher, 1998). It has been suggested that variation in gO may confer differences in cell tropism (Jarvis and Nelson, 2007).

Phylogenetic analysis of UL74 (gO) and UL115 (gL) sequences from 40 low-passage clinical samples by Rasmussen *et al.* (2002) led to the description of four major genotypes for both genes, although UL115 sequences varied by less than 2%. In contrast, UL74 sequences varied by as much as 46%. No association between genotype and patient type was found.

Rasmussen *et al.* (2002a) also analysed gH (UL75), gL (UL115) and gO (UL74) in 84 samples by RFLP and found evidence for genetic linkage between gH1 and gO1. This could be due to their close proximity on the genome (they are adjacent genes) or their functional interaction in the gCIII complex. However, this does not explain the lack of linkage observed between other gH and gO genotypes, nor does it explain the absence of linkage between gL and gO or gL and gH. It should also be noted that they grouped UL74 sequences into only four genotypes based on RFLP analysis rather than sequencing, therefore their suggestions of linkage should be treated with care.

Mattick *et al.* (2004) sequenced the hypervariable N-terminal region of gO (UL74) in 50 unpassaged clinical isolates and described four major groups (gO1, gO2, gO3 and gO4) with some further division into subtypes: gO1a, gO1b and gO1c, and gO2a and gO2b. Intergenotypic variation was high, whereas within genotypes the sequences were highly conserved. Phylogenetic analysis of intergenotypic alignments suggested that some residues were under positive selection and, when branch lengths were allowed to vary, positive selection was detected close to the base of the tree. Based on this finding, the authors postulated that gO sequences were under positive selection early in their history and that if the sequences have since been under purifying selection, fixation of synonymous changes could mask other positively selected residues. As these workers had previously genotyped gN (UL73) and gB (UL55) in some of these samples, they investigated potential linkage between these genes and found strong evidence for linkage between gN and gO genotypes, which could reflect their proximity on the HCMV genome (they are only 28 nt apart on the genome). However, gH and gO are also adjacent genes, but they are over 400 nt apart and the lack of linkage between gH and gO genotypes may be a consequence increased recombination due to this increased distance. Mixed infections were detected, albeit at low frequencies (four of the 50 samples).

1.6.8 UL123

UL123 is a MIE gene that encodes the IE1 protein, which is essential for viral replication *in vivo* and *in vitro* and is a transactivator that positively autoregulates IE1/IE2 and U3 expression (Mocarski *et al.*, 1996). Sequencing of the fourth exon of the IE1 gene (MIE exon 4) in seven samples from immunocompromised patients and the laboratory strains AD169 and Towne identified five aa sequence changes in two of the seven patients (Brytting *et al.*, 1992). The sequences were conserved over time both within patients and when passaged in cell culture. A mixed infection was identified in a single patient.

A larger study by Retiere *et al.* (1998) sequenced MIE exon 4 in 25 clinical isolates, and phylogenetic analysis of these strains plus AD169 and Towne revealed three groups. These workers also sequenced gB in these isolates and found no evidence for linkage disequilibrium between MIE exon 4 genotypes and gB (UL55) genotypes, nor did they find any evidence for linkage between genotype and pathogenesis.

A more recent study investigated sequences of MIE exon 4 (as well as gB (UL55) and UL97) in HCMV strains from six immunocompromised patients, plus AD169 and Towne (Mousavi-Jazi *et al.*, 2000). The aim was to evaluate a potential link between genotype and replication rate by monitoring HCMV gene expression and the production of infectious virus. No evidence for a connection between MIE exon 4 genotype (or gB (UL55) or UL97 genotype) and viral replication was found, and all sequences fell into groups 1 and 3 as determined by Retiere *et al.* (1998).

1.6.9 UL144

UL144 encodes a tumour necrosis factor (TNF) α-like receptor that may play a role in HCMV virulence by facilitating evasion of the immune system (Benedict *et al.,* 1999). This finding led to interest in UL144 as a potential marker of pathogenesis and to the discovery that UL144 is highly variable, with 21% nucleotide and aa sequence divergence between 45 clinical isolates (Lurain *et al.,* 1999). Phylogenetic analysis revealed three major genotypes. Investigation of UL144 sequences in 62 passaged samples from congenitally infected neonates (23 from living neonates and 39 autopsy samples from ten foetuses) again revealed three major genotypes (A, B and C).

Two recombinant subtypes A/B and A/C (i.e. an A type sequence in the N-terminal part of the gene and a B- or C-type sequence in the C-terminal part of the gene) were also identified (Arav-Boger *et al.,* 2002). Variation was concentrated in the N-terminal region and mixed infections were detected in eight of the ten autopsied samples. Some evidence for an association between UL144 genotype and disease outcome was found, suggesting that infection with the most commonly detected genotype, genotype B, conferred a more favourable prognosis.

A number of studies have investigated UL144 sequences in various patient types, and all have described three major genotypes A, B and C, which were detected in 97% of samples. The recombinant subtypes (A/B and A/C) were detected in only 3% of samples. These subtypes may be due to a PCR artefact (Section 1.7), and duplicate experiments are required to confirm their presence. All genotypes were distributed amongst seropositive infants and adults and among symptomatic and asymptomatic foetuses, suggesting a lack of association between UL144 genotype and clinical disease (Bale *et al.,* 2001; Mao *et al.,* 2007; Picone *et al.,* 2005).

1.6.10 The *a* sequence

The *a* sequence is a small repeat sequence (~600 bp) found at the HCMV genome termini and also in inverted orientation between IR_L and IR_S (see Figure 1.1 and Figure 1.3). It contains two conserved motifs, *pac*1 and *pac*2, both of which are required for cleavage and packaging of DNA during replication (Spaete and Mocarski, 1985). Phylogenetic analysis of the *a* sequence in 39 low-passage HCMV isolates revealed six distinct groups. The largest group contained 16 isolates with >95% nucleotide identity. The same study examined potential linkage between *a* sequence groups and gB (UL55) genotypes and found no evidence for linkage disequilibrium (Bale *et al.,* 2001). A more recent study investigated the *a* sequence in 74 HCMV clinical isolates from 60 Japanese infants and children (collected from 1983 to 2003), and phylogenetic analysis revealed five groups (Tanaka *et al.,* 2005).

1.6.11 Short tandem repeats

A short tandem repeat (STR) or microsatellite is a DNA sequence motif of 1-6 nucleotides that is repeated. They are found in eukaryotes and some prokaryotes and tend to be hotspots of length mutation, possibly due to replication slippage (Field and Wills, 1998).

The HCMV genome contains at least 24 STRs, and examination of their sequences in ten passaged clinical isolates plus AD169 and Towne revealed variation between strains (Davis *et al.,* 1999). Many STRs are located in non-coding regions of the genome, and variation in length, as well as point mutations, occur. For the majority, two or three variants were detected, but ten variants were detected for one region.

A later study by Walker *et al.* (2001) examined the sequences of ten STRs in 44 clinical isolates plus AD169 and Towne, in order to examine the utility of STRs for HCMV strain characterisation. These workers developed a PCR-based assay, which utilised primers specific for ten STRs, to compare the STR patterns obtained for each, and found they could accurately differentiate between HCMV strains. They found STRs to be highly variable (one to 12 variants) and suggested that multiplex STR analysis rather than multiple gene genotyping, which is time consuming, could be used for strain comparison.

Table 1.2. A Selection of Studies of Variation in HCMV Genes

Variable gene or region	*Protein*	*Number of genotypes (% aa sequence divergence)*	*References*
UL1	gpUL1 (RL11 family)	3	Sekulin *et al.*, 2007
UL4	gp48 (RL11 family)	4	Alderete *et al.*, 1999; Bar *et al.*, 2001; Sekulin *et al.*, 2007
UL6	gpUL6 (RL11 family)	4	Sekulin *et al.*, 2007
UL7	gpUL7 (RL11 family)	3	Sekulin *et al.*, 2007
UL10	gpUL10 (RL11 family)	3	Sekulin *et al.*, 2007
UL11	gpUL11 (RL11 family)	3 (43%)	Davison *et al.*, 2003a; Hitomi *et al.*, 1997
UL37	gpUL37 and gpUL37$_M$	5 (28%), based on exon 3	Hayajneh *et al.*, 2001, 2001a
UL55	gB	gB1-gB4 (9.5%) (3 rare gB5-gB7)	Chou and Dennison, 1991; Meyer-König *et al.*, 1998; Pignatelli *et al.*, 2004; Shepp *et al.*, 1998; Trincado *et al.*, 2000
UL73	gN	gN1-gN4c (7 including subtypes) (50%)	Mattick *et al.*, 2004, Pignatelli *et al.*, 2001, 2002, 2003
UL74	gO	gO1-gO5 (7 including subtypes), 20% (40% at N-terminus)	Mattick *et al.*, 2004, Paterson *et al.*, 2002; Rasmussen *et al.*, 2003, Stanton *et al.*, 2005
UL75	gH	21% at nt level, only 5% at aa level	Rasmussen *et al.*, 2003
UL123	IE1	Specifically within exon 3 (28%)	Brytting et al., 1992; Mousavi-Jazi *et al.*, 2000
UL139	gpUL139	3 (5 subtypes)	Qi *et al.*, 2006
UL144	gpUL144 (TNF-α-like receptor)	gA,gB,gC (subtypes gAC and gAB) or g1,g2,g3 (22%)	Arav-Boger *et al.*, 2002; Bale *et al.*, 2001; Coaquette *et al.*, 2004; Lurain *et al.*, 1999; Picone *et al.*, 2005
UL146	vCXCL-1	14 (G1-G14)	Arav-Boger *et al.*, 2005, 2006, 2006a; Dolan e*t al.*, 2004; Lurain *et al.*, 2006; Stanton *et al.*, 2005; He *et al.*, 2006
UL147	vCXCL-2		Arav-Boger *et al.*, 2005; He *et al.*, 2006; Lurain *et al.*, 2006
STR	N/A	24 STRs,1-15 genotypes	Davis *et al.*, 1999; Picone *et al.*, 2005; Walker *et al.*, 2001
a sequence	N/A	6 (65%)	Bale *et al.*, 2001; Tanaka *et al.*, 2005

STRs have been described in other viruses and it has been suggested that variation in these elements may affect virulence. Variation in the length of trinucleotide repeats at the haemagglutinin glycoprotein cleavage site of avian influenza virus has been associated with enhanced virulence (Perdue *et al.,* 1997). Analysis of seven STRs in 47 HCMV clinical isolates from congenitally infected infants and immunocompromised individuals revealed a greater number of alleles for these STRs (up to 15 variants) than previously reported, although this may reflect the larger sample number. The same study found no evidence for association between STR alleles and clinical disease (Picone *et al.,* 2005a).

1.7 Mixed HCMV infections

The occurrence of more than one strain in an HCMV infection (a mixed infection) complicates genotyping studies and consequent attempts to identify potential associations between genotype and disease. It also has important connotations for vaccine design, particularly since pre-existing immunity to one strain offers only partial protection against reinfection with another strain (Boppana *et al.,* 2001). With MCMV, mixed infections have been reported in 23-67% of free-living mice, and experiments using two MCMV strains found that laboratory mice could be infected simultaneously or successively with more than one strain, even in the presence of MCMV-specific antibody and CTL responses (Gorman *et al.,* 2006).

Mixed infections may also have important implications for transplant recipients, where there have been reports that such infections may result in increased viral load, HCMV disease and subsequent rejection of the transplant (Coaquette *et al.,* 2004; Gerna *et al.,* 1992; Puchhammer-Stöckl *et al.,* 2006).

The proportion of mixed infections is likely to be underestimated as a consequence of methods employed for diagnosis and genotyping. Specifically, isolation of virus from clinical samples by cell culture may result in selection of certain virus variants and the loss of others.

Additionally, analysis of a single hypervariable locus is insufficient for assessment of mixed infections, as, for example, patients with a single gB (UL55) genotype were found to have more than one gN (UL73) genotype (Puchhammer-

Stöckl *et al.,* 2006). This is not surprising, given the lack of linkage between hypervariable loci (Rasmussen *et al.,* 2003). This is also likely to be a reflection of the higher level of variation observed for gN (UL73) when compared to gB (UL55), thus making it a more sensitive marker.

Furthermore, PCR conditions need to be optimised, as the presence of multiple related sequences in a reaction can result in a recombination artefact, which is thought to be a consequence of incomplete primer extension during elongation steps. Incomplete extension can occur if short extension times are employed or the DNA secondary structure interferes with DNA polymerase binding, causing the polymerase to 'fall off' (Judo *et al.,* 1998; Qiu *et al.,* 2001). The incompletely extended primer can anneal to a different, partially complementary template in a subsequent cycle and undergo extension to produce a recombinant or 'chimera' (Judo *et al.,* 1998). In regards to HCMV, identification of recombinant molecules by PCR following co-infection experiments suggested a high frequency of recombination between HCMV strains infecting the same cell (Sevilla-Reyes, 2007). However, this was shown to be due to the PCR artefact described above.

It may be that coinfection with more than one strain facilitates complementation and results in enhanced virus fitness. Attenuated MCMV strains have been shown to benefit from coinfection and can complement each other *in vivo* via *trans*-complementation (Čičin-Šain *et al.,* 2005).

1.8 Recombination in HCMV

Recombination produces genetic diversity and can accelerate genetic divergence as it produces new alleles or genotypes (Mayr, 2001). Homologous recombination and non-homologous (or illegitimate) recombination have both been described in herpesviruses (Dohner *et al.,* 1988; Henderson *et al.,* 1990; Nishiyama *et al.,* 1991). Homologous recombination is recombination between two pieces of DNA containing sequence homology, whereas non-homologous recombination occurs in the absence of sequence homology.

Homologous and non-homologous recombination are likely to have played a role in HCMV evolution. For example, host genes are thought to have been captured through recombination between the viral genome and host sequences and subsequent duplication of some of these genes has produced gene families such as the CXC-chemokines (UL146 and UL147). Duplication events are also likely to be the result of recombination (Arav-Boger *et al.*, 2005; Davison *et al.*, 2003a; Sahagun-Ruiz *et al.*, 2004).

The absence of linkage disequilibrium between hypervariable loci in HCMV is consistent with recombination having played an important evolutionary role (Rasmussen *et al.*, 2003). Despite this, there has been no evidence of recombination within genes producing different genotypes at hypervariable loci, with the exception of rare forms of gB and one gO genotype (Mattick *et al.*, 2004).

1.9 UL146

The hypervariable gene UL146 encodes a CXC (or α) chemokine designated vCXC-1. UL146 in strain Toledo encodes a fully functional chemokine that produces chemotaxis, calcium mobilisation and neutrophil degranulation. vCXC-1 binds to human CXCR2, and its activities are comparable to those of human chemokines IL-8 and gro-α (Penfold *et al.*, 1999). UL146 is thought to promote virus dissemination through this ability to attract monocytes to the initial site of infection. Phylogenetic analysis of UL146 sequences in 17 unpassaged clinical isolates (urine, whole blood and tissue samples) identified 14 genotypes (Dolan *et al.*, 2004). Variation is distributed throughout the gene. Only the characteristic CXC chemokine motif and four additional residues are completely conserved between genotypes. UL146 sequences are stable over time, both *in vitro* when passaged in fibroblasts, and *in vivo* by sequencing isolates from the same patient over a period of several years (Stanton *et al.*, 2005; Lurain *et al.*, 2006).

Numerous studies (some published during the course of this thesis work) have investigated whether UL146 genotype correlates with HCMV disease (Stanton *et al.*, 2005; Hassan-Walker *et al.*, 2004; Lurain *et al.*, 2006; He *et al.*, 2006). All of these studies utilised relatively small sample sizes (11–50 patients) and all, with the exception of one study (He *et al.*, 2006), examined only immunocompromised individuals (neonates, AIDS patients and transplant recipients). Connections between clinical outcome and UL146 genotype were not detected.

He *et al.* (2006) investigated HCMV strains circulating in 25 infants or young children and all but two showed symptoms of HCMV-associated disease. This group did not use the genotype nomenclature established previously (Dolan *et al.*, 2004), and instead divided UL146 sequences into three major groups (G1, G2, G3) with some divided into subgroups (G1A, G1B, G2A and G2B), making a total of five groups. No significant connection between UL146 genotype and clinical outcome was found, although it was noted that two asymptomatic individuals contained the same genotype (G2B). He *et al.* (2006) conceded the limitations associated with using such a small sample size. Based on these groupings, evidence for linkage between UL146 and UL144 genotypes was detected. However, this finding could have been compromised by the use of a smaller number of genotypes (three) than those reported previously (14, Dolan *et al.*, 2004).

Lurain *et al.* (2006) sequenced UL144, UL146, UL147, UL147A and the intergenic region (between UL146 and UL147) in 50 clinical isolates. UL146, UL147 and the intergenic region were highly variable. All UL146 sequences grouped into the 14 genotypes described previously (Dolan *et al.*, 2004). All UL147 sequences and intergenic sequences also clustered into 14 groups. No evidence for linkage disequilibrium between UL146 genotypes and UL144 genotypes was found. UL146 was expressed as a true L gene on a single transcript (3.7 kb) that includes UL147, UL147A and UL132 (Lurain *et al.*, 2006).

HCMV encodes additional cytokines such as UL147 (a putative CXC-chemokine adjacent to UL146 for which no functional data has been reported) and UL111A (vIL-10). An alternatively spliced UL111A transcript (UL111.5A) is expressed during latency, which is speculated as encoding an IL-10 homologue that may prevent immune recognition (Jenkins *et al.*, 2004). CCMV contains two homologues of HCMV UL146, CCMV UL146 and UL146A (Davison *et al.*, 2003). CCMV UL146 is a functional CXC-chemokine that can induce calcium mobilisation and chemotaxis, although it binds CXCR2 with lower affinity than Toledo-encoded vCXCL-1 (Miller-Kittrell *et al.*, 2007). RhCMV also encodes a number of UL146-related genes (Alcendor *et al.*, 1993; Penfold *et al.*, 2003).

1.10 UL139

UL139 is predicted to encode a type I membrane glycoprotein. As observed for other hypervariable glycoproteins, variation is concentrated at the 5’-end of the gene encoding the putative ectodomain (Dolan *et al.,* 2004). During the course of this thesis work, Qi *et al.* (2006) reported the sequences of UL139 in 19 low passage clinical isolates (fewer than ten passages in human embryonic lung fibroblasts) and seven urine samples from 26 HCMV-positive infants. All sequences fell into three major groups (G1, G2 and G3) with two divided into subgroups, making a total of six genotypes (G1A, G1B, G1C, G2A, G2B and G3). Variable numbers of predicted N-linked glycosylation, casein kinase II phosphorylation and N-myristoylation sites were reported. Mixed infections (three genotypes) were detected in three patients (~12%). No association between UL139 genotype and disease was found, although the authors conceded that the sample size was too small to make definitive conclusions in this regard. A region of sequence similarity between all variants of the UL139 protein and CD24 was noted (SETTTGTSSNSSQST). This region is rich in serine and threonine residues that could be potentially O-glycosylated. This region is located just downstream of the predicted signal sequence cleavage site and upstream of the highly conserved C-terminal region.

CD24 is a cellular glycosyl phosphatidylinositol-linked glycoprotein that is a signal transducer involved in B cell activation (Fisher *et al.,* 1990). Additional roles for CD24 in apoptosis and cell adhesion have also been suggested, and more recently in regulating the responsiveness of a chemokine receptor, CXCR4 (Smith *et al.,* 2006; Schabath *et al.,* 2006). It has been suggested that CD24’s role as a ligand for P-selectin could help tumour cells exit from the bloodstream and hence promote metastasis (Kristiansen *et al.,* 2004). Variation in glycosylation has been observed in CD24 and has been linked to differences in cell and tissue specificity (Goris *et al.,* 2006; Poncet *et al.,* 1996). This similarity to CD24 is intriguing as it suggests a possible immunomodulatory role or even a role in tissue tropism for UL139. However, no expression or functional data have been published for UL139.

1.11 Specific Objectives of the study

UL146 and UL139 are two of the most hypervariable HCMV genes. Both probably have roles in regulation of the immune response, and indeed, Toledo-encoded UL146 has been shown to encode a functional chemokine (Penfold *et al.,* 1999). To date, 14 UL146 genotypes (Dolan *et al.,* 2004) and, during the course of this thesis work, six UL139 genotypes (Qi *et al.,* 2006) have been described. The initial aim of this study was to characterize UL146 and UL139 sequences in a much larger panel of clinical isolates than examined previously, from a range of distinct geographical locations and clinical settings.

Specific foci of the study were as follows:

- The total number of UL146 and UL139 genotypes in circulation and their frequencies of occurrence
- The geographical distribution of genotypes
- The modes of evolution that may have given rise to the different genotypes
- To understand the effects of *in vitro* and *in vivo* passage and the generation of hypervariation
- The potential genetic linkage between UL146 and UL139 genotypes
- The frequency of mixed infections
- The potential structural differences between the proteins ecoded by UL146 genotypes
- The transcription of UL146 and UL139
- The basic characterisation of the UL139 protein

UL146 and UL139 are located in the region at the right end of U_L that is absent from laboratory passaged stocks of AD169 (Cha *et al.,* 1996). However, a stock of AD169 (AD169*var*UC) included in the genotyping study was found to contain both genes. The sequence of the entire region at the right end of the U_L that is absent from normal AD169 stocks was determined.

2 Materials and Methods

Materials

2.1 Viruses

The collection of 184 HCMV virus strains used in the genotyping study (Chapter 3) consisted of 179 anonymised clinical samples from disparate geographical locations plus five commonly used laboratory strains (Davis, Merlin, TB40/E, Toledo and Towne). Some strains were derived by collaborators as routine diagnostic specimens grown in human fibroblast cell culture (maximum of 5 passages, indicated in Table 3.1). They were kindly supplied as DNA extracted from body tissues, urine, saliva or infected cells by collaborators as detailed in Table 2.1. Full details of the collection are summarized in Table 3.I. DNA extraction, PCR, cloning and sequencing of the Hungarian and Dutch samples (37) was performed by Ida Kovács (University of Szeged). Australian, Gambian, Hungarian, Dutch and Chinese samples were extracted using the Nucleospin Tissue kit (Macherey-Nagel) according to the manufacturers instructions for "Purification of CMV DNA from urine". Scottish samples were extracted using a robot and the Qiagen DNA extraction Kit (with the exception of the CSF samples which were extracted manually using Qiagen colums). German and South African samples were extracted using the Qiagen blood extraction kit, in a dedicated room in which no herpesvirus experiments were performed. All extractions were performed by experienced operators working to category II standards, taking strict precautions to prevent genomic contamination, which included negative controls.

HCMV strain Merlin (previously called isolate 742;Tomasec *et al.,* 2000; Dolan *et al.,* 2004) was kindly provided by Prof. G. W. G. Wilkinson (Cardiff University) and was used in all transcript mapping studies.

AD169*var*UC was generously provided by Prof. N. Lurain (University of Chicago) via Prof. P. Ghazal (University of Edinburgh).

2.2 Cells and cell culture media

Phosphate buffered saline (PBS)	8 g/l NaCl 0.2 g/l KCl 1.44 g Na_2HPO_4 0.24 g KH_2PO_4
Versene	0.2 g/l EDTA in PBS 0.002% (w/v) phenol red
Giemsa strain	Sigma-Aldrich
10 X trypsin solution	Invitrogen
Polyfect transfection reagent	Qiagen
Tetrachloroethylene	Acros Organics
10 X citric saline solution	100.6 g/l KCl 44.12 g/l sodium citrate
MEM non-essential amino acids 100 X, without L-glutamine	Biosera Ltd
L-glutamine (200 nM)	Invitrogen
Penicillin-streptomycin (10000 U/ml)	Invitrogen
Foetal calf serum (FCS)	Invitrogen

2.3 Oligonucleotides

Custom DNA oligonucleotides were designed for PCR, RACE (Table 2.2) and QPCR. QPCR genotypic primers are shown in Table 3.12. PCR and sequencing primers for the right end of U_L in AD169*var*UC are shown in Table 2.3. All primers were manufactured by Sigma-Aldrich and provided as lyophilised solids, which were resuspended in distilled water to a final concentration of 100 μM.

Table 2.1. HCMV Strain Sources

Samples	*Viral samples*	*DNA extraction*
A1-A18	William D. Rawlinson and Gillian M. Scott, Prince of Wales Hospital, Sydney, Australia	Derrick Dargan, MRC Virology Unit
C1-C10	Paul K. Chan, Prince of Wales Hospital, Shatin, Hong Kong, China	Derrick Dargan, MRC Virology Unit
D1-D7	Thomas F. Schulz and Khaled R. Alkharsah, Hannover Medical School, Hannover, Germany	Charles Cunningham, MRC Virology Unit
E1-E12	Vincent C. Emery, Division of Infection and Immunity, Royal Free and University College Medical School, London, England. Paul A. Moss, University of Birmingham, Birmingham, England	Derrick Dargan, MRC Virology Unit
G1-G18	Steve Kaye, MRC Laboratories, Banjul, The Gambia	Derrick Dargan, MRC Virology Unit
H1-H30	Rozalia Pusztai and Ida J. Kovács, University of Szeged, Hungary	Ida J. Kovács, University of Szeged
I1-I7	Giuseppe Gerna, IRCCS Policlinico, San Matteo, Italy	Derrick Dargan and Charles Cunningham, MRC Virology Unit
N1-N6	Rozalia Pusztai and Ida J. Kovács, University of Szeged, Hungary	Ida J. Kovács, University of Szeged
S1-S45	William F. Carman and Bassam B. Ismaeil, Gartnavel General Hospital, Glasgow, Scotland Colin C. Geddes, Western Infirmary, Glasgow, Scotland	Bassam B. Ismaeil, Gartnavel General Hospital
U1-U5	Alistair McGregor, University of Minnesota, Minneapolis, USA	Alistair McGregor, University of Minnesota. Derrick Dargan, MRC Virology Unit
W2-W9	Gavin W. G. Wilkinson, Cardiff University, Cardiff, Wales	Gavin W. G. Wilkinson, Cardiff University. Charles Cunningham, MRC Virology Unit
Z1-Z15	Martin Dedicoat, Ngwelezane Hospital, Empangeni, KwaZulu-Natal, South Africa Thomas F. Schulz and Khaled R. Alkharsah, Hannover Medical School, Hannover, Germany	Khaled R. Alkharsah, Hannover Medical School

2.4 PCR, QPCR and SMART RACE PCR

Advantage 2 polymerase	
10 x PCR buffer	
Advantage UltraPure PCR	
dNTP mix	BD Clontech

SYBR® Green PCR master mix Applied Biosystems

(containing SYBR® Green 1 Dye, AmpliTaq Gold® DNA Polymerase LD, dNTPs with dUTP/dTTP blend, ROX and optimized buffer components)

Table 2.2. Primers used for PCR, Sequencing and RACE

Gene/Plasmid	*Primer*	*Sequence (5'-3')*	*Location*[a]
UL146	AB4	TAGACACTACGTCGTAAATG	180494-180513
UL146	A162	TGTAGAATTAGTCTAGATTCCTGA	181524-181501
UL146	UL146-4A	GCTTGCGCGTTAGGATTGAGACAC	180571-180594
UL146	UL146-3A	ATACCGGATATTACGAATT	181341-181323
UL139	AB1	GTCATTGTGAAAGTGACGTCTCAG	186389-186412
UL139	AB2	ATCTACTGTAAACCCTCTGCTCTG	187148-187125
UL139	UL140-3A	GCGGCATTGGTGTACGCGTG	187058-187078
UL139	UL140-11A	GTGGAAATTTTTACGTCATT	186572-186553
pGemT	F21	ACGTTGTAAAACGACGGCCAG	N/A
pGemT	R21	CACACAGGAAACAGCTATGAC	N/A
UL139	5'UL139RACE	CAGCAGCTGGACACTTTACGTACTAGCC	186607-186634
UL139	3'UL139RACE	CTGCTGGTACCACTAACACGACTACACC	186790-186763
UL140	3'UL140RACE	TCGGCTTCATCGTTACGCTAC	186186-186166
UL141	3'UL141RACE	GTGTTGGTCGCCGAGGGAGAG	185250-185230
UL146	5'UL146RACE	CACCTGTTATCGTTGCGTTTGTCTAGCC	181009-181036
UL146	3'UL146RACE	GTGCATGGAACGGAATTACGCTG	181235-181213
UL139	NthUL139FWD	ATGCTGTGGATATTAGTTTTATTTG	186878-186853
UL139	NthUL139REV	TAAAGGTGGAGGCGGAGCCACT	186484-186462
UL146	NthUL146FWD	ATGCGATTAATTTTTGGTGCG	181292-181271
UL146	NthUL146REV	GGATCATCCAGACTTCCTTATT	180952-180930
N/A	SMART II A	AAGCAGTGGTATCAACGCAGAGTACCGGG	N/A[*]
N/A	5' RACE CDS A	$(T)_{25}$VN (N=A,C,G,T; V= A,C,G)	N/A[*]
N/A	3' RACE CDS A	AAGCAGTGGTATCAACGCAGAGTAC$(T)_{30}$VN	N/A[*]
N/A	Long UPM	CTAATACGACTCACTATAGGGCAA GCAGTGGTATCAACGCAGGT	N/A[*]
N/A	Short UPM	CTAATACGACTCACTATAGGGC	N/A
pAL942	PMV100f	GTGAACCGTCAGATCGCC	N/A
pAL942	PMV100r	AGTACGGTTTCACAGGCG	N/A
UL54	UL54fwd	ACGGCCAAACCATGTCATGACTCA	81670-81693
UL54	UL54rev	GTCATGTTCGACGGTCAGACG	81778-81758

[a] With reference to RefSeq accession NC_006273.2 (HCMV strain Merlin). Where the second coordinate is larger than first, primers are in the rightward orientation on genome. Where the reverse is the case, primers are leftward oriented.

[*] Supplied with SMART RACE™ cDNA Amplification kit

Table 2.3: PCR and Sequencing primers for right end of AD169

PCR Product	Primer	Sequence (5'-3')	Genome location[a]
RL5A	C3	CAGAGTTATACTATAGTC	4928-4945
	C5	GTTGACCTAGTTAGATTT	5493-5476
RL13	A7	GATACATGCGTCGTATGCCGCCAC	8943-8966
	A10	ATTCCAAACCGGATACGCTACATA	13096-13073
UL11	A13	GTATGGAGGTCACTGTCAGAGTAG	17517-17540
	A16	GGACAGCTGGTACGTCGCTCCTTG	19490-19467
UL73	A76	CATGCAAACGAATTGCGCGTCCAG	106158-106181
	A77	CATGCACGACTCGGACGACGTCCT	109057-109034
UL131A	UL132-11	GCCATGCAACCCGTCTCGCT	177903-177884
	UL132-4	GTCATGCGGTTTGGAATACG	177085-177104
UL148	UL146-8A	CTGAGACGTCATGCTGGTAG	212553-212534
	UL132-10A	CAGCAACCCGAACGCGACCA	178612-178631
UL122	A155	ACGGTACATAGTTACCCTCTCGAC	169963-169986
A155	A156	TAGAGTTCTTTACCAAGAACTCAG	173170-173147
A156	A121A	CCTGTGGAAGGTAGATTACGACAG	170082-170059
	A157	TGATCAATGTGCGTGAGCACCTTG	172996-173019
	UL122-3	TGTCTTCTTATCACCATCAG	172195-172176
	UL122-6	GGTTTAATAATCACCTTGAA	171947-171966
	UL122-7	GAACAGGGTGAAGAAGTCGA	171810-171791
	UL122-10	GCACACCCAACGTGCAGACT	171383-171364
	UL122-11	TCCGCCACTGCTGCATTTCA	171577-171596
	UL122-12	TAGCGTGGCATTGATGGTCA	170244-170263
	UL122-14A	GGCGCTCTCAACCTGTGCCT	170919-170900
	UL122-15	TGTGCTCCATGAGGAAGGGA	171090-171109
	UL122-16	GGACACTGTGTCTGTCAAGTCTGA	172473-172450
	UL122-19A	CCACACGTTAATACTGTCAC	170622-170641
	UL122-20	GGGAGACTTAGAATCTCTTG	170459-170440
	UL123-1C	ACAGGCGTGACACGTTTATTGAGT	172217-172240
	UL123-3	ACTAGGAGAGCAGACTCTCA	172704-172723
	UL123-11	GGCTGAGAACAGTGATCAGGAAGA	172548-172525
	UL123-12	TATGGATATCCTCACTACAT	172989-172970
UL132	A159	CTCATATCGTCTGTCACCTATATC	175854-175877
A159	A160	GTTTACTCCTCGTGTTGCAAGCAC	178752-178729
A160	A158	TATTGAAAATGTCGCCGATGTGAG	176050-176027
	A161	TGAGAACCTCGTCGGGAACCGCTG	178622-178645
	UL123-2	CAACTACAATCCGTAAGTCT	176534-176515
	UL123-10	CGCGGCACACATCCAGCCGTTTGT	176212-176235
	UL128-1	ATCCCGCGAATCTCAGCCGT	176595-176614
	UL130-1A	CTGTAGTCCCGGAAGACGTG	177069-177088
	UL132-1	TGGGACTCATGACGCGCGGT	176935-176954
	UL132-2	AATGTTGCGAATTCATAAACGTCA	176859-176836
	UL132-4	GTCATGCGGTTTGGAATACG	177476-177495
	UL132-6	CCACATACTTGTAACGGGTT	177987-178006
	UL132-8	ACGAACGACGTGTCCAAGTT	178463-178482
	UL132-11	ATAGTGCGATGGCGTTTGTG	190758-190739
	UL132-12	TGCGACGACAGCCGCGTGGT	189222-189241
	UL132-13	TTGTATAGCAGCACACGCCT	188814-188795
UL146	A161	TGAGAACCTCGTCGGGAACCGCTG	178622-178645
A161	A162	TGTAGAATTAGTCTAGATTCCTGA	181524-181501
A162	A163	AATTCGTAATATCCGGTATTCCCG	181323-181346
A160	HYP-2A	GTGCAATGCATACTGTCCCAGTCG	179892-179915
	UL132-3	ACCCGTGGTGGAAAATGTTG	179741-179722
	UL132-5	ACAGATTCATCGTGCAGTAC	179241-179222
	UL132-7	TTCAGCTTCATAGCGGTACT	178782-178763
	UL132-9	GCGACGCAGCGTCCAGTTCA	179474-179493
	UL132-10A	CAGCAACCCGAACGCGACCA	179002-179021
	UL146-2	GCGTACCGCAAATCACTAGG	180183-180202
	UL146-4A	GCTTGCGCGTTAGGATTGAGACAC	180571-180594
	UL146-7B	GCGAGCGAAAGCTGCAATCGTCAG	180653-180630
	UL146-8A	CTGAGACGTCATGCTGGTAG	180163-180144

UL144	A163	AATTCGTAATATCCGGTATTCCCG	181323-181346
A162	UL144-1	TGTCTCCCTGGGCCACTCGG	184598-184579
A162	Hyp-3A	CTAGTGTTACATCGATACAGTGCC	181761-181738
UL144-1	UL144-3A	ATTCGGATACTTTGTGTCAT	184297-184278
	UL144-5A	ACTACCTGCATAGAAAGACT	183768-183749
	UL144-6B	AGGCTAGAGTATGACGACC	182331-182349
	UL144-7A	CACCTTACAGCATATGAGCA	183362-183343
	UL144-8A	CATAACTTCACTAACCCGCA	182831-182850
	UL144-9	GGTAACTATCGTAAGTCGGTAGGC	184422-184445
	UL144-10	TGAGATACGCGATGAATGTT	183989-184008
	UL144-12A	GTTTTCCGAACTTTTATACA	183056-183075
	UL144-13	TGTATAAAAGTTCGGAAAAC	183075-183056
	UL144-14	TTCTTCCGGTAGGAGGCATG	182753-182734
	UL144-15	TGCCAACAGTGTTGCTCAAT	182253-182234
UL140	UL144-9	GGTAACTATCGTAAGTCGGTAGGC	184422-184445
UL144-9	UL138-1A	CTGATCCGCTGTTGCGAGCTGTAC	187812-187789
UL138-1A	UL133-8	CATGGCTACGGTGGTGAACTGCGT	187468-187491
	UL140-1	TGACATTCTCTGCTCGATCT	187394-187375
	UL140-2	TATAGAAGTAGTTGCGTTGA	184776-184795
	UL140-3A	GTGGAAATTTTTACGTCATT	187077-187058
	UL140-4	TCTCGGCCCACATCTTTTCG	185277-185296
	UL140-5	CGTCACTTTCACAATGACGT	186406-186387
	UL140-6	CTGATGAAGCTGCCAAGAGT	185714-185733
	UL140-7	GTCGTACTAACAGCGTGTCA	186009-185990
	UL140-8	GCGTCGCACGGTGGTCACCA	185520-185501
	UL140-9	GCCACTTGGAATTTCTCGCA	185073-185054
	UL140-10A	GAGAAAGAAAAGTAGCGTAA	186155-186174
	UL140-11A	GCGGCATTGGTGTACGCGTG	186553-186572
	UL140-12	CAGAGCAGAGGGTTTACAGT	187125-187144
	UL144-1	TGTCTCCCTGGGCCACTCGG	184598-184579
	UL144-9	GGTAACTATCGTAAGTCGGTAGGC	184422-184445
UL136	UL133-8	CATGGCTACGGTGGTGAACTGCGT	187468-187491
UL133-8	A164	CAGGCCCTTCCCGAAAACGCCGAC	189107-189084
A164	UL133-10C	ACGAACGACGTGTCCAAGTT	178463-178482
UL138-1A	UL133-13	TTGTATAGCAGCACACGCCT	188814-188795
	UL133-14	CATCACGCCGATGATGGGTA	187903-187922
	UL133-21	TCTGCCGCTCGTGGTGCCGA	188392-188411
	UL133-22	TATCTCCCGCTACGTAAGAG	188049-188030
	UL133-23	AGACATGCTCCACGATCTAT	188462-188443
UL133	UL133-10C	CCTTCATGACGCTCTGCACCGCCT	188872-188895
UL133-10C	UL133-7	TTCAGCTTCATAGCGGTACT	178782-178763
UL133-7	UL133-11A	ATAGTGCGATGGCGTTTGTG	190758-190739
A164	UL133-12	TGCGACGACAGCCGCGTGGT	189222-189241
UL150-1	UL133-15	GCGTAAGAACCTGAGCACGC	189194-189175
	UL133-16	GACATCGGAACCCAAACCGA	190207-190188
	UL133-17	ACGACGTCTTCTTTCGGA	189573-189590
	UL133-18	ACCGGACTGGCTTCCCTGGT	189733-189714
	UL133-19	CGGGTGGCATCTGCGGCATG	190035-190054
	UL133-20	CACGCTGAACAGCAGCGGCT	190397-190416
UL150	UL150-1	CTAGTAACACTCGTCCGACACTTC	190818-190841
UL150-1	A165	GAACGCCGTGCACCACAAACTCTG	193757-193734
A165	A166	GAACGTCGTCCTCCCCTTCTTCAC	193582-193605
	UL133-2	CAGCGCCCAGGCGATCTCGCGCTC	191638-191615
	UL133-7	TTCAGCTTCATAGCGGTACT	178782-178763
	UL150-3	GCAGGATAGCGGTTAAGGAT	191237-191256
	UL150-5	GCAGGATAGCGGTTAAGGAT	191237-191256
	UL150-6	AACCCACGTTAACCGACCGT	191756-191775
	UL150-7	TTCGTCCACGGTCTCCGAGA	193400-193381
	UL150-8	CGACAACGCCATCAGGAGAT	192205-192224
	UL150-9A	GGATGGCCGTCCGTCGAAGC	191322-191303
	UL150-10A	CAGATAGTTCCACGGACAG	193651-193669
	UL150-11B	CCGCGACTCCTCCAGGTTG	192135-192117
	UL150-12B	GCTGCGTAAAGTACATCAG	192293-192275
	UL150-13	CGGAGCTCGTTGGCGCGGAA	192645-192664
	UL150-14	AACAGCGACGCGACTTTGGG	193179-193198

REND	A166A	AGCAGCGAGCTACGCAGACGGAAT	193248-193271
A166A	UL150-21	CGCTAYTCTTTATTAACGTC	194319-194300
UL150-21	IRS-1A	CGTTGGAGAATTGGTGGGATCGGT	193904-193927
A165	IRS-7	GACGGCGAATGCAGCAGACGGTGT	194061-194084
A166	JC	TGGGCCATGTGTGGTGGCAG	194161-194180
UL150-6	UL150-4B	CTCGCTGTTGCGCCACCTCTT	193838-193818
UL150-10A	UL150-16	TCACAGCGACATGTTGCTTCGTC	194241-194219
	Adrend1	AACGACACAGGCAAGGAC	[b] 726-709/189672-189689
	Adrend2	AACTAGTCGCCGTCCACAC	[b] 88-70
	Adrend3	GATCCACTGGAGCGCACAG	[b] 190462-190447/229646-229661
RENDL	JC	TGGGCCATGTGTGGTGGCAG	194161-194180
JC	A167	GCCCAGCGCCAGGTACAGTCCGTC	196608-196585
A167	A122	ATGGCCCAGCGCAACGGCATGTCG	195976-195999/234759-234736
	JA	ACCCAGCACACGGCCCGGAATGGA	195645-195668/235090-235067
	JB	TCCATTCCGGGCCGTGTGCTGGGT	195668-195645/235067-235090
LENDL	A167	GCCCAGCGCCAGGTACAGTCCGTC	196608-196585
A12, A167	A2	TTTCGGCGTGAAGTTGGACGGCGT	2204-2181
JA	A1	ACAGGCTTTCGCGCACACGATTCC	1883-1906
JB	IRS-5	CCCACATGCACCAGCAGTCGGCGT	1784-1761
	Adlend1*	CGACATGCCGTTGCGCTGG	195999-195981/234736-234754
	Adlend2*	CTCAGCCACGGTTCACAATC	2152-2171
	Adlend3*	GACGCGGCGCGAACAGC	[b]865-850/189533-189548
	Adlend4*	CCAACACCGTCCCGCACA	[b] 818-801/189580-189597
	Adlend5*	CGACATGCCGTTGCGCTGGG	195999-195980/234736-234755

[a] With reference to RefSeq accession NC_006273.2 (HCMV strain Merlin)

[b] With reference to Genbank accession BK_000394.2 (HCMV strain AD169)

Note forward primers, where second coordinate is larger than first are in rightward orientation on HCMV genome, and those where the reverse is the case are leftward oriented. All primers with the exception of those marked with an asterix (*) were designed by Andrew Davison and ordered by Charles Cunningham. The PCR primers used to produce each PCR fragment are coloured blue (as is the PCR fragment). Primers used to amplify a product were also used to sequence a product. Additional sequence primers are shown in black. Where the sequence of a primer has already been given, the name of the primer used for sequencing is below the PCR product. UL146 and UL139 were both amplified by nested PCR and sequenced using primers listed in Table 2.2 as described in Section 2.11.

2.5 Whole genome amplification

REPLI-g whole genome amplification kit-

REPLI-g DNA polymerase

4 X REPLI-g buffer

Control gDNA template (10 ng/μl)

Solution B (stop solution)

1 X PBS

1 M DTT Qiagen

2.6 Agarose gel electrophoresis

Agarose	Sigma-Aldrich
10 X TBE	109 g/l Tris 55 g/l boric acid 9.3 g/l EDTA
DF dyes	37.2 g/l EDTA 100 g/l Ficoll 400 5 X TBE 1% (w/v) bromophenol blue
DNA marker (2 log ladder)	New England BioLabs
Ethidium bromide	10 mg/ml aqueous solution Sigma-Aldrich
Geneclean turbo spin kit- Turbo salt solution Wash concentrate Catch tubes and spin filters	Q-biogene
PureLink Quick Gel Extraction kit- Gel solubilization buffer (GS1) Wash Buffer (W9) Catch tubes and spin filters	Invitrogen
HiDi formamide	Applied Biosystems

2.7 Plasmid preparation

One shot TOP10 *Escherichia coli* strain K12 competent cells (Invitrogen).

Genotype: F^- *mcr*A Δ(*mrr-hd*RMS-*mcr*BC) ϕ80*lac*ZM15 Δ*lac*X74 *deo*R recA1 araD139 Δ (ara-leu) 7697 galU galK rpsL (Str^R) endA1 nupG.

SW102 [SW101 ΔgalK (DH10B [lc1857 (cro-bioA)<>Tet] galK+ gal490)] cells containing pAL942 was used for production of recombinant adenoviruses (Warming *et al.,* 2005) and was kindly supplied by Dr. Richard Stanton, Cardiff University.

pAL942 contains the human adenovirus 5 genome (AdEasy system, Invitrogen) in a bacterial artificial chromosome (BAC) with the HCMV MIE promoter (containing tet operators), HCMV IE polyA site, amp, SacB, LacZ, and a C-terminal streptavidin (strep) tag. It was kindly supplied collaboratively by Dr. Richard Stanton, Cardiff University.

pGEM-T Vector System I- pGEM-T vector T4 DNA ligase 2 X ligation buffer Control insert DNA	Promega
L-broth	10 g/l NaCl 5 g/l yeast extract 10 g/l tryptone peptone Becton Dickinson
L-broth agar	1.5% (w/v) agar in L-broth
SOC medium	Invitrogen
Ampicillin (made up to 100 μg/μl)	Melford Laboratories
Chloramphenicol (made up to 100 μg/μl)	Sigma-Aldrich
X-gal (5-bromo-4-chloro-3-indoyl-β- D-galactopyranoside in N, N’-dimethyl formamide	40 mg/ml Invitrogen

IPTG	30 mg/ml
(isopropylthio-β-galactoside)	Invitrogen
2YT broth	5 g/l NaCl
	1% (w/v) bactopeptone
	10 g/l yeast extract
Restriction endonucleases	
and buffers	NEB/Roche
QIAprep miniprep kit-	
Buffer P1, RNase A	
Buffer P2, buffer N3	
Collection tubes	Qiagen
QIAfilter Plasmid Maxi kit-	
Qiagen-tip 500	
Qiafilter Maxi cartridges	
Buffer P1, RNase A	
Buffer P2, buffer P3	
Buffer QC, buffer QBT	
Buffer QF	Qiagen
Electroporation cuvettes	BioRad

2.8 RNA preparation and extraction

Cycloheximide	Sigma-Aldrich
Phosphonacetic acid (PAA)	Sigma-Aldrich
$((HO)_2P(O)CH_2CO_2H)$	
TRI reagent	Sigma-Aldrich
Chloroform	Sigma-Aldrich

Isopropanol	Sigma-Aldrich

2.9 Northern blotting

10 X MOPS (MOPS [3-(N-morpholino) propanesulfonic acid])	41.2 g/l
Loading buffer	500 μl/ml deionized formamide 166 μl/ml 37% formaldehyde 100 μl/ml 10 X MOPS 100 μl/ml 99% RNase-free glycerol
RNA molecular weight marker I, digoxigenin-labelled	Roche
20 X SSC	88.2 g/l tri-sodium citrate 174 g/l NaCl
Maleic acid buffer pH 7.5 with NaOH pellets	11.61 g/l maleic acid 8.76 g/l NaCl
Detection buffer pH 9.5 with concentrated HCl	12.11 g/l Tris 5.84 g/l NaCl
Nylon membrane, positively charged	Roche
3MM Paper	Whatman
DIG-Northern starter kit- 5 X Labelling mix 5 X Transcription buffer SP6 RNA polymerase, 20 U/μl T7 RNA polymerase, 20 U/μl Anti-DIG-alkaline phosphatase antibody DNase I, RNase-free	

CDP-*Star* chemiluminescent substrate	
Actin RNA probe, DIG-labelled	
DIG Easy Hyb granules	
10 X Blocking solution	Roche
Chemiluminescent film	Roche
Diethylpyrocarbonate (DEPC)	
97%, density 1.12 g/ml	Sigma

2.10 Western blotting

Resolving gel buffer	181.5 g/l Tris
(RGB)	4 g/l SDS
Spacer gel buffer	59 g/l Tris
(SGB)	4 g/l SDS
Running buffer	6.32 g/l Tris
	4 g/l glycine
	1 g/l SDS
Transfer buffer	3.025 g/l Tris
	14.4 g/l glycine
	0.3 ml/l concentrated HCl
	200 ml/l methanol
Acrylamide	
(N, N’-methylene-bis acrylamide)	BioRad
Ammonium persulphate (APS)	BioRad
TEMED (N, N, N’, N’-	
tetramethylethylenediamine)	Sigma-Aldrich
Rainbow markers RPN756	Amersham

Blocking solution	5% (w/v) Marvel milk 10% (v/v) FCS 10% (v/v) glycerol (99% v/v) in PBS
Hybond (H^+) nylon membrane	Amersham
ECL detection reagents	GE healthcare
Photographic film	Kodak Ltd

2.10.1 Antibodies

Anti-FLAG M2 antibody	Stratagene
Anti-actin antibody	Sigma-Aldrich
Goat anti-rabbit IgG HRP	BioRad
Goat anti-mouse IgG HRP	BioRad

Methods

2.11 Polymerase chain reaction

For the genotyping of UL146 and UL139 in clinical samples, both genes were amplified separately by single or nested PCR, using primers in conserved regions (Table 2.2). Single (and first) round PCR of UL146 using AB4 and A162 generated a product of approximately 1 kbp, and second round PCR using UL146-4A and UL146-3A yielded an 800 bp product. Single (and first) round PCR of UL139 using AB1 and AB2 generated an 800 bp product, and nested PCR using UL140-3A and UL140-11A yielded a 500 bp product.

PCR was performed as follows:

10 X Advantage 2 Buffer	5 µl
dNTPs (10 µM each)	1 µl
Forward primer (10 µM)	1 µl
Reverse primer (10 µM)	1 µl
Sterile distilled water	40 µl
Advantage 2 DNA polymerase	1 µl (1 U)
Template (extracted DNA)	1 µl

The conditions for amplification were 95°C for 2 min followed by 35 cycles of 95°C for 2 minutes, 60°C (this temperature was adjusted for the melting temperatures of primers used) for 30 sec and 68°C for 1 min/kbp of product length. Second round PCR utilized 1 µl of first round PCR products as template amplified under the same conditions. PCR reactions were set up in a dedicated, PCR product-free room. Approximately one-third of the samples were tested in triplicate to examine the reproducibility of results.

2.12 Agarose gel electrophoresis

PCR products were electrophoresed through a 1% (w/v) agarose (1 X TBE) gel at 100 V for approximately 2 h. Gels were stained with ethidium bromide (0.5

μg/ml) for 30 min and photographed under short wavelength UV light using the GelDoc system (BioRad).

2.13 Recovery of DNA fragments and cloning

2.13.1 Purification of PCR products

PCR products were excised from an agarose gel under long wavelength UV light and purified using a Geneclean turbo kit according to the manufacturer’s instructions (Q-biogene). Smaller PCR products (<100 bp) were purified using a PureLink gel extraction kit according to the manufacturer’s instructions (Invitrogen). Products were eluted in 50 μl of sterile distilled water and stored at -20°C.

2.13.2 Ligations

Purified PCR products were ligated into the pGEM-T vector as follows:

2 X T4 DNA ligase buffer	2.5 μl
pGEM-T vector	0.25 μl
Sterile distilled water	0.25 μl
T4 DNA ligase	0.5 μl
DNA insert (0.1-0.5μg)	1.5 μl

Reactions were incubated at 4°C overnight.

2.13.3 Bacterial transformations

2.13.3.1 Chemical transformation

Competent cells were allowed to thaw on ice for 15 min. Ligation mixture (1 μl) was added to a 17 μl aliquot of cells and incubated on ice for 30 min. Cells were heat-shocked at 42°C for 30 s and then chilled on ice for 5 min. SOC medium (250 μl) was added to each sample and the samples were incubated at 37°C with shaking (180 rpm) for 1 h. IPTG (10 μl) and X-gal (20 μl) were added to each tube and mixed, and the solution was spread on an L-agar plate (containing ampicillin at 100 μg/ml) using a sterile plastic spreader. The plates were allowed to dry at room temperature (RT) for 1 h and then incubated upside-down at 37°C overnight.

2.13.3.2 Preparation of electrocompetent cells

A single colony of pAL942 was inoculated into 5 ml of LB (containing 100 μg/ml ampicillin and 40 μg/ml chloramphenicol) and incubated 32°C with shaking (180 rpm) overnight.

An aliquot of this overnight culture (1 ml) was used to inoculate 50 ml of LB (containing 100 μg/ml ampicillin) and incubated at 32°C with shaking (180 rpm) until the culture had reached an OD_{600} of 0.6. The culture was split into 2 X 50 ml Falcon tubes (i.e. 25 ml in each tube). One tube was incubated at 42°C for 15 min to induce the λ red proteins and the other tube was incubated at 32°C for 15 min as a negative control. Both tubes were then chilled on ice for 15 min. The cells were centrifuged at 2,200 x g for 5 min at 4°C. The supernatant was discarded and the pellet was resuspended in 1 ml of ice-cold water by gently swirling the tube. Once the pellet was resuspended, 25 ml of ice-cold water was added and the tube was centrifuged as before. This 'wash' step was repeated. Following centrifugation, the supernatant was discarded and the cells were resuspended in the small amount of liquid remaining (~100 μl).

2.13.3.3 Electroporation

Cells, cuvettes, DNA and tubes were chilled on ice. An aliquot of cells (25 μl) was placed in a tube to which 4 μl DNA (~0.5 μg) was added. This was then transferred to a 0.1 cm cuvette and incubated on ice for 5 min. The cells were electroporated at 2.5 kV, 25 μF and 400 Ω. Samples were recovered in 5 ml volume of LB at 32°C with shaking (180 rpm) for 4 h. A 50 μl volume of each sample (and a number of tenfold dilutions, 1:10, 1:100 and 1:1000) were plated on L-amp [without NaCl, with sucrose (6% w/v), chloramphenicol (40 μg/ml), IPTG (10 μl) and X-gal (20 μl)] plates and incubated at 32°C for 36 h.

2.14 Whole genome amplification

The genomic DNA template (2 μl) was denatured by the addition of 2.5 μl of denaturation buffer [Buffer D1 (made up fresh, 10 μl solution A (containing KOH and EDTA), 70 μl sterile distilled water)]. The sample was mixed and incubated at RT for 3 min. A 5 μl aliquot of neutralisation buffer [Buffer N1 (made up fresh, 16 μl solution B (containing KCl, Tris-HCl and HCl), 14 μl sterile distilled water)] was then added and the sample was mixed again, this is the denatured genomic template used in the next step.

The REPLI-g polymerase was allowed to thaw on ice and reactions were set up as follows:

Denatured genomic DNA template	9.5 μl
Sterile distilled water	27 μl
4 X REPLI-g buffer	12.5 μl
REPLI-g polymerase	0.5 μl

Samples were incubated at 30°C for 8 h.

The REPLI-g polymerase was then inactivated and reaction stopped by incubating the samples at 65°C for 3 min.

2.15 Plasmid DNA preparation

2.15.1 Miniprep plasmid purification

Bacterial colonies grown on L-agar plates were inoculated into 1.5 ml of 2YT broth (containing appropriate antibiotic) and incubated at 37°C with shaking (180 rpm) overnight. For the production of recombinant adenoviruses (RADs) bacterial colonies grown on L-agar plates [with sucrose (6% w/v) for SW102 cells] were inoculated into 1.5 ml of LB (containing appropriate antibiotic) and incubated at 32°C with shaking (180 rpm) overnight.

Plasmid purification was carried out using the QIAspin miniprep kit according to manufacturer's instructions (Qiagen).

2.15.2 Maxiprep plasmid purification

Bacterial colonies were inoculated into 5 ml of LB and incubated at 32°C with shaking (180 rpm) for ~8 h. The entire 5 ml culture was used to inoculate 500 ml of LB and incubated at 32°C with shaking (180 rpm) overnight.

Large-scale plasmid purification was performed using the maxiprep kit according to manufacturer’s instructions (Qiagen).

2.16 DNA sequencing

Sequencing reactions were performed in 96-well plates as follows:

5 X ABI buffer	1.75 μl
BigDyes	0.5 μl
Sterile distilled water	1.75 μl
Primer (1.6 μM)	2 μl
DNA	4 μl

The sequencing programme consisted of a denaturation step of 95°C for 2 min followed by 30 cycles of 94°C for 10 sec, 50°C for 5 sec and 60°C for 4 min. Following this, the resultant DNA was ethanol-precipitated by the addition to each well of 62 μl of ethanol mix (containing 10 ml 100% ethanol, 3 ml sterile distilled water, 400 μl of 3 M sodium acetate). The plate was agitated to mix and centrifuged at 6,000 x g for 30 min at 4° C. The supernatant was removed by inverting the plate. The pelleted DNA was then washed by the addition of 150 μl of 70% (v/v) ethanol to each well and centrifugation at 6,000 x g for 30 min at 4°C. This wash step was repeated before one final brief centrifugation (plate inverted) at 440 x g to dispel all remaining ethanol. The pellets were allowed to dry at RT before being resuspended in HiDi formamide for loading onto a capillary sequencer (ABI 3730). Electrophoresis of sequencing reactions and data collection were performed by an external service provider at the BHF Glasgow Cardiovascular Research Centre (Genomics Laboratory).

2.16.1 DNA sequence analysis

Sequence chromatograms were viewed and processed using Editview (Applied Biosystems) and Pregap4 and Gap4 (Staden *et al.,* 2000). All nucleotide and derived aa sequences were aligned using ClustalW (Thompson *et al.,* 1994) and Mafft (Katoh *et al.,* 2005), and alignments were corrected manually using Bioedit. Both ClustalW and Mafft are Unix-based programs. Nucleotide sequences were then realigned using the aa alignments as a template. ClustalW uses the progressive method to align sequences, firstly aligning the two most closely related sequences in a set by the neighbour-joining method using position specific gap penalties. It then successively aligns the next most closely related sequence to the alignment produced in the previous step. A phylogenetic tree is produced at the same time. The final alignment is constructed by combining all alignments produced, in the order specified by the tree, maintaining all gaps. One limitation of this method is that it depends heavily on the quality of the initial alignment. Mafft is an alternative multiple alignment program that offers three alternative strategies, one based on the progressive method and two based on an iterative method (with refinements available such as the weighted sum-of-pairs score and consistency scores). The iterative method works similarly to the progressive method but it repeatedly realigns the initial sequences as well as adding new sequences to the growing alignment.

The Phylip package (Felsenstein, 1989) and Mega 4.0 (Tamura *et al.,* 2007) were used for the generation of phylogenetic trees based on the neighbour-joining method. Phylip (*phy*logeny *i*nference *p*ackage) is Unix-based and infers phylogenies (evolutionary trees) by parsimony, distance-matrix and likelihood methods. Mega (molecular evolutionary genetic analysis) is a multifunctional PC program that can be used for the analysis of an alignment. Mega can infer phylogenetic trees and test evolutionary hypotheses using a number of alternative methods.

Frequencies of nonsynonymous and synonymous differences per site (dN and dS, respectively) and degree of sequence variability (nucleotide and aa) were investigated using Swaap 1.0.1 (Pride, 2004), MEGA4.0 and PAML 3.15 (Yang, 1997). Swaap (sliding windows alignment analysis program) allows basic analyses of an alignment such as sequence identity, nucleotide composition, transitions/transversions and dN/dS over a whole gene or a sliding window.

Signal peptide and transmembrane sequences were predicted using Phobius (Kall *et al.*, 2004, 2007).

2.17 3D homology modelling

Homology models were built using the Molecular operating environment (MOE) protein modelling and 3D bioinformatics software (Molecular Operating Environment 2003.02, The Chemical Computing Group Inc., 2003.).

2.18 Quantitative PCR using SYBR green

Quantitative PCR (QPCR) was performed in 96-well plates using an Applied Biosystems 7500 Fast Real-Time PCR instrument as follows:

2 X SYBR Green PCR master mix	10 μl
Forward primer (18 μM)	1 μl
Reverse primer (18 μM)	1 μl
Sterile distilled water	7 μl
DNA	1 μl

The conditions for amplification were 95°C for 10 min followed by 35 cycles of 95°C for 15 sec, 60°C (this temperature was adjusted for melting temperature of the primers used) for 1 min and 72°C for 25 sec (with the plate read during the exponential phase) followed by dissociation analysis from 65-95°C, with the plate read at every 0.2°C increment and then held at 4ºC. PCRs were set up in a dedicated, PCR product-free room.

2.19 Cell culture

2.19.1 HFFF-2 cells

Human foetal foreskin fibroblast (HFFF-2) cells (initially supplied by Dr. Derrick Dargan, MRC Virology Unit) were used for growth of HCMV strain Merlin and for studies of both HCMV and RADs gene expression. HFFF-2 cells were grown in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% (v/v) FCS, 1% (v/v) non-essential amino acids, 1% (v/v) L-glutamine supplements and 1% (v/v) penicillin-streptomycin at 37°C in a humidified atmosphere of 95% (v/v) air and 5% (v/v) CO_2. Confluent HFFF-2 cells were harvested by discarding the medium, washing the monolayer with versene (twice), and then detaching the cells from the flask using a 1:20 dilution of trypsin in versene. The cells were resuspended in 10 ml of fresh medium and seeded into 175 cm^2 flasks at a ratio of 1:2 (i.e. one flask split into two new flasks), with 40 ml of fresh medium added per flask.

2.19.2 HEK 293 cells

Human embryonic kidney (HEK) 293 cells (initially supplied by Dr. Katarina Baluchova, MRC Virology Unit) were used for large-scale production and titration of RAD stocks. HEK 293 cells were grown in DMEM supplemented with 10% (v/v) FCS, 1% (v/v) L-glutamine and 1% (v/v) penicillin-streptomycin at 37°C in a humidified atmosphere of 95% (v/v) air and 5% (v/v) CO_2. Confluent HEK 293 cells were harvested by discarding the medium, washing the monolayer with versene (twice), and then detaching cells from flask using a 1:20 dilution of 10 X citric saline solution in versene. The cells were then resuspended in 10 ml of fresh medium and seeded into 175 cm^2 flasks at a ratio of 1:2, with 40 ml of fresh medium added per flask.

2.20 Preparation of virus stocks

2.20.1 HCMV

HFFF-2 cells were seeded into an 80 cm^2 flask until they had reached 80% confluency. The medium was discarded and the cells were then infected with HCMV strain Merlin at multiplicity of infection (m.o.i.) of 0.1 p.f.u./cell in 4 ml of supplemented DMEM. The cells were incubated at 37°C for 1 h, after which a further 20 ml of medium was added to the flask and the cells were incubated at 37°C. Cytopathic effect (CPE) was monitored by examining the infected cells under a light microscope each day. Once the cells had reached 80% CPE the infected cells were harvested and used to infect one 175 cm^2 flask of HFFF-2 cells. When this flask had reached 80% CPE, the infected cells were harvested and used to infect four 175 cm^2 flasks. When they had reached 80% CPE, the infected cells were harvested and used to infect 14 roller bottles, which were then incubated at 37°C until they had reached 100% CPE (usually 10-15 days post infection (p.i.)). The medium containing infected cells was decanted and cells were pelleted by centrifugation at 1,000 x g at 4°C for 10 min. The supernatant was decanted into large Sorvall tubes and centrifuged at 9,700 x g at 4°C for 20 min. Cell pellets from both spins were combined and stored at -70°C. The supernatant was collected, aliquoted and stored at -70°C prior to titration.

2.20.2 Recombinant adenoviruses

2.20.2.1 Transfection of HEK 293 cells with recombinant adenoviruses

HEK 293 cells were seeded into 25 cm^2 flasks (10^6 cells/flask) and incubated at 37°C overnight. RAD DNA (4 μg) was made up to 150 μl with DMEM (non-supplemented). Polyfect transfection reagent (40 μl) was added to the DNA, vortexed gently to mix and incubated at RT for 10 min. The monolayer was washed gently with PBS complete, and 3 ml of fresh medium was added. The DNA/polyfect mixture was made up to 1 ml with medium and added to the flask. The cells were incubated at 37°C overnight. Medium was discarded, fresh medium was added, and the cells were incubated at 37°C until they had reached 100% CPE.

2.20.2.2 Harvesting recombinant adenoviruses

Cells were detached from the flask by gently tapping the side of the flask and transferred with the medium to a 15 ml Falcon tube. The tubes were centrifuged at 470 x g for 10 min at 4°C. The supernatant was discarded and the pellet was resuspended in 1 ml of versene by gently tapping the tube. An equal volume of tetrachloroethylene was added and mixed well. The tube was centrifuged at 470 x g for 10 min at 4°C. The top layer was removed carefully and transferred to a fresh microcentrifuge tube. Half of this was removed and stored at -70°C. The other half was made up to 5 ml with medium to be used for large-scale production of RADs. This was used to infect ten 175 cm^2 flasks, which were then incubated at 37°C until they had reached 100% CPE.

Cells were detached from all ten flasks by gently tapping side of flasks and transferred to 50 ml Falcon tubes. The tubes were centrifuged at 470 x g for 10 min at 4°C. The supernatant was discarded and the pellets were resuspended in 1 ml of PBS. All pellets were then combined and centrifuged again. The final pellet was resuspended in 5 ml of PBS to which an equal volume of tetrachloroethylene was added. This was centrifuged at 470 x g for 10 min at 4°C and the top layer was carefully transferred to a tube before virus aliquots were prepared.

2.21 Titration of virus

2.21.1 HCMV

HCMV titres were determined by plaque-assay on HFFF-2 monolayers grown in 24-well tissue culture plates (5.10^4 cells/well). Virus stocks were serially diluted, and 100 μl of a range of dilutions (10^{-3}, 10^{-4}, 5.10^{-5},10^{-5}, 5.10^{-6} and 10^{-6}) were plated on cell monolayers in triplicate. Virus was allowed to adsorb to the cells for 1 h at 37 ºC, with gentle rocking every 15 min. Following virus adsorption, the cell monolayers were overlaid with 1 ml of supplemented medium and incubated at 37 °C until visible plaques were observed (10-12 days p.i.). The medium was removed and 1 ml of Giemsa stain was added, and incubated at RT for 6 h. After staining, the fixed cell layers were rinsed thoroughly and the plaques were counted using a light microscope.

2.21.2 Recombinant adenoviruses

RAD titres were determined by the $TCID_{50}$ assay (the 50 percent effective tissue culture infective dose). HEK 293 cells were grown in 96-well flat-bottomed tissue culture plates (2.10^6 cells/plate, 100 μl/well). Virus stocks were serially diluted, and 100 μl of ten-fold dilutions from 10^{-2} to 10^{-10} were plated on the HEK 293 monolayers in replicates of ten per plate. The plates were incubated at 37°C until visible plaques were observed (7-10 days p.i.).

2.22 Preparation of HCMV RNA

HFFF-2 cells were seeded into 24-well plates at 60-80% confluency (10^6 cells/well) and incubated at 37°C for 16 h. HCMV strain Merlin was used for all infections. Immediate early (IE), early (E) and late (L) RNAs were prepared as follows. Mock-infected (MI) RNA was prepared similarly.

2.22.1 IE RNA

The medium was discarded and 2 ml of fresh medium (containing 200 μg/ml cycloheximide) was added to each well. The plate was incubated at 37°C for 1 h. The medium was discarded and virus was added at an m.o.i. of 3 p.f.u./cell to 20 of the 24 wells (fresh medium was added to the remaining four wells for mock infected samples). The plate was incubated at 37°C for 1 h. The cells were washed three times with 2 ml of fresh medium (containing 200 μg/ml cycloheximide) per well. Fresh medium (2 ml containing 200 μg/ml cycloheximide) was added to each well, and the cells were incubated at 37°C for 24 h. The medium was discarded, and the cells were washed three times with 2 ml of fresh medium (containing 200 μg/ml cycloheximide) per well. TRI reagent (200 μl) was added to each well and the cells were homogenised by pipetting up and down and scraping the monolayer using the pipette tip. The homogenised samples were transferred to microcentrifuge tubes and stored at -70°C.

2.22.2 E RNA

The medium was discarded and 2 ml of fresh medium [containing 200 μg/ml phosphonoacetic acid (PAA)] was added to each well. The plate was incubated at 37°C for 1 h. The medium was discarded and virus was added at an m.o.i. of 3 p.f.u./cell to 20 of the 24 wells (fresh medium was added to the remaining 4 wells for mock infected samples). The plate was incubated at 37°C for 1 h. The cells were washed three times with 2 ml of fresh medium (containing 200 μg/ml PAA) per well. Fresh medium (2 ml containing 200 μg/ml PAA) was added to each well, and the cells were incubated at 37°C for 48 h. The medium was discarded, and the cells were washed three times with 2 ml of fresh medium (containing 200 μg/ml PAA) per well. TRI reagent (200 μl) was added to each well and the cells were homogenised by pipetting up and down and scraping the monolayer using the pipette tip. The homogenised samples were transferred to microcentrifuge tubes and stored at -70°C.

2.22.3 L RNA

The medium was discarded and 2 ml of fresh medium was added to each well. The plate was incubated at 37°C for 1 h. The medium was discarded and virus was added at an m.o.i. of 3 p.f.u./cell to 20 of the 24 wells (fresh medium was added to the remaining 4 wells for mock infected samples). The plate was incubated at 37°C for 1 h. The cells were washed three times with 2 ml of fresh medium per well. Fresh medium (2 ml) was added to each well, and the cells were incubated at 37°C for 72 h. The medium was discarded, and the cells were washed three times with 2 ml of fresh medium per well. TRI reagent (200 μl) was added to each well and the cells were homogenised by pipetting up and down and scraping the monolayer using the pipette tip. The homogenised samples were transferred to a microcentrifuge tube and stored at –70°C.

2.22.4 Preparation of total cellular RNA

Homogenised cell samples were allowed to thaw and incubated at RT for 5 min. Chloroform (200 μl) was added to each sample and mixed vigorously for 15 sec. The samples were incubated at RT for 10 min. They were then centrifuged at 14,000 x g for 30 min at 4°C.

The aqueous (upper) phase was transferred to a fresh microcentrifuge tube. Isopropanol (500 μl) was added to each sample and mixed. The samples were incubated at RT for 10 min. They were then centrifuged at 14,000 x g for 20 min at 4°C.

The supernatant was removed, paying careful attention not to dislodge the pelleted RNA. An aliquot of 1 ml of 70% (v/v) ethanol was added to each tube to wash the pellet, and the tube was vortexed and centrifuged at 14,000 x g for 5 min at 4°C.

The supernatant was removed, paying careful attention not to dislodge the pelleted RNA, and the pellet was air-dried for 10 min. The RNA pellet was dissolved in 10 μl of RNAse-free water (preheated to 55°C).

2.22.5 Determination of RNA yield

RNA was quantified using a spectrophotometer. To assess the quality of cellular RNA, 1 μg of total RNA was electrophoresed on 1% (w/v) agarose (1 X TBE) gel at 100 V for 2 h, and the gel was stained with ethidium bromide (0.5 μg/ml) for 30 min. RNA was visualised over a short-wave UV transilluminator, and photographed using the BioRad Gel Doc system. High quality RNA is expected to exhibit a ratio of 2:1 for the 28S and 18S rRNAs. Samples that produced smeared 28S and 18S bands, or that deviated from the expected ratio of 2:1, were discarded.

2.23 Northern blotting

To avoid RNase contamination, disposable plastic-ware was used whenever possible and all plastic was autoclaved twice. Glassware was autoclaved and baked twice in a dry oven; gel tanks and other re-useable plastic-ware were washed with RNaseZap (SIGMA), rinsed in distilled water and allowed to air-dry. DEPC-treated water was used throughout, where DEPC was added at a concentration of 0.1% to sterile distilled water, incubated at 37°C overnight (to remove any RNAses), before being autoclaved to remove any residual DEPC.

2.23.1 Agarose gel electrophoresis of RNA

A formaldehyde-agarose gel was prepared by dissolving 1.5 g of agarose in 141.9 ml of 1 X MOPS, then cooling to 55°C before the addition of 8.1 ml of formaldehyde. Loading buffer was prepared fresh, added to total RNA (1 μg) and RNA ladder (at ratio 2:1, v/v) and mixed. Samples and ladder were incubated at 65°C for 10 min and chilled on ice for 5 min. The gel tank was filled with 1 X MOPS buffer, samples were loaded, and electrophoresis was carried out at 40 V for 6 h. The gel was allowed to cool for 15 min before manipulation. The gel was washed twice in 20 X SSC for 15 min (by shaking gently at 150 rpm) to remove formaldehyde.

2.23.2 Transfer of RNA to membrane

RNA was transferred from the gel to a nylon membrane by capillary blotting. Nylon membrane was cut to size and pre-soaked in water (5 min) and then 20 X SSC for 20 min. 3MM paper was also cut to size and pre-soaked in 20 X SSC. The gel was placed on 3MM paper that descended into a reservoir of 20 X SSC. The nylon membrane was placed on top of the gel, followed by 3 MM paper and a stack of dry paper towels with a weight on top.

RNA was allowed to transfer overnight. The membrane was then rinsed gently in water, dried for 20 min and RNA crosslinked to the membrane using a Stratagene UV crosslinker (12,000 J/cm^2).

2.23.3 Preparation of RNA probes

RNA probes were generated using purified PCR products ligated into the multiple cloning site of the plasmid p-GemT, which is flanked by Sp6 and T7 RNA polymerase promoter sites. UL146 was amplified using the primers NthUL146FWD and NthUL146REV and UL139 was amplified using NthUL139FWD and NthUL139REV (Table 2.2). A number of independent clones were prepared and sequenced. Those plasmids that contained the gene in the correct orientation were linearised using a restriction enzyme (SalI or NdeI) that leaves a 5'-overhang. Following restriction digestion, the linearised plasmid was purified using the Geneclean turbo spin kit (Q-biogene) according to the manufacturer's instructions. Linearised plasmid DNA was eluted in 50 μl of sterile distilled water and quantified using a spectrophotometer. Transcriptional labelling of probes was performed as follows:

5 X Labelling mix	4 μl
5 X Transcription buffer	4 μl
Sterile distilled water	5 μl
T7 RNA polymerase	2 μl
DNA (1 μg)	5 μl

Reactions were set up on ice, vortexed gently to mix and centrifuged briefly before being incubated at 42°C for 1 h. DNase I (2 μl) was added to each tube and the tubes were incubated at 37°C for 15 min to remove template DNA. A 2 μl aliquot of 0.2 M EDTA (pH 8) was added to each tube and the tubes were mixed thoroughly. Probes were stored at -20°C.

2.23.4 Nucleic acid hybridisation

Probes were quantified by preparing ten-fold serial dilutions from 10^{-2} to 10^{-6}, and spotting the dilutions onto a nylon membrane. The probe spots were cross-linked to the membrane using a Stratagene UV crosslinker (12,000 J/cm^2). Bound probe was then detected as outlined in Section 2.23.5 (using one-fifth volume described for all buffers). A DIG-labelled actin probe was used as a positive control.

The blot was placed in a hybridization tube and 15 ml of pre-warmed DIG Easy Hyb buffer was added. The blot was incubated with rotation at 68°C for 1 h. The amount of probe required (100ng/ml of Hyb buffer) was transferred to a microcentrifuge tube and made up to 50 μl with RNase-free water. The tube was incubated at 68°C for 10 min and then chilled on ice. The contents were then added to 3 ml of fresh, pre-warmed DIG Easy Hyb buffer. The 15 ml of DIG Easy Hyb buffer was discarded, and 3 ml of fresh buffer containing the probe was added to each hybridisation tube. This was incubated with rotation at 68°C overnight.

The blot was washed (twice) in 20 ml of low stringency buffer at RT with shaking. The blot was then washed (twice) in 40 ml of pre-warmed high stringency buffer with rotation at 68°C for 15 min.

2.23.5 Detection

The membrane was transferred to a tray containing 100 ml of wash buffer and incubated at RT with shaking for 5 min. Wash buffer was discarded and 100 ml of blocking solution was added. The blot was then incubated at RT with shaking for 30 min. The blocking solution was discarded and the blot was incubated in 20 ml of antibody solution at RT with shaking for 30 min. The membrane was washed (twice) in 100 ml of wash buffer at RT with shaking for 15 min. The membrane was equilibrated in 20 ml of detection buffer for 3 min. The detection buffer was drained and the membrane was placed on a plastic sheet. CDP-*star* (about 4 drops) was added to each membrane and spread evenly, and the membrane was then incubated at RT for 3 min. Excess CPD-*star* was removed and the plastic sheet was sealed (ensuring no bubbles were present).

The sealed plastic envelope containing the blot was exposed to chemiluminescent film for varying lengths of time to obtain a range of images.

2.24 5'- and 3'-RACE

SMART RACE produces full-length cDNAs by reverse transcription using the SMART II oligonucleotide and murine leukemia virus reverse transcriptase (MMLV RT). When the MMLV RT reaches the end of the RNA template it acts as a terminal transferase and adds 3–5 residues (predominantly dC) to the 3' end of the first strand cDNA. The SMART II oligo (oligonucleotide) contains a terminal stretch of G residues (3-5), which anneal to the dC-tail, and this serves as an extended template for RT. MMLV RT then switches from the mRNA template to the SMART II oligo and generates a complete cDNA copy of the original RNA, which contains the additional sequence of the SMART II oligo. This cDNA is then used as a template for 5'- and 3'-RACE using a gene specific primer and a universal primer that binds to the SMART II sequence.

2.24.1 Synthesis of 5'- and 3'-RACE-ready cDNA

cDNAs were synthesised from total cellular RNAs using a SMART™ RACE cDNA amplification kit (BD Clontech) according to the manufacturer's instructions, as follows:

5'-RACE-ready cDNA		3'-RACE-ready cDNA	
5'-CDS Primer	1 μl	3'-CDS Primer	1 μl
SMART II oligo	1 μl	Sterile water	3 μl
Sterile water	2 μl	Total RNA (1 μg)	1 μl
Total RNA (1 μg)	1 μl		

The tubes were vortexed gently to mix and centrifuged briefly. The samples were incubated at 70°C for 2 min, and then chilled on ice for 2 min before being centrifuged briefly. The following was then added to each tube:

5 X First-strand buffer	2 μl
Dithiothreitol (DTT) 20 μM	1 μl
dNTP mix (10 μM)	1 μl
Powerscript reverse transcriptase	1 μl

The tubes were vortexed gently to mix and centrifuged briefly. The samples were incubated at 42°C for 1.5 h. The tubes were then made up to 100 μl with TE buffer and incubated at 72°C for 7 min to stop the reaction.

2.24.2 5'- and 3'-RACE PCR

RACE PCR reactions were set up as follows:

5'-/3'-RACE-ready cDNA	2.5 μl
10 X UPM	5 μl
Gene specific primer (10 μM)	1 μl
Sterile distilled Water	34.5 μl
10 X Advantage 2 buffer	5 μl
dNTP (10 μM)	1 μl
Advantage 2 polymerase	1 μl

Thermocycling conditions were as follows: 5 cycles of 94°C for 30 sec, 72°C for 3 min; followed by 5 cycles of 94°C for 30 sec, 70°C for 30 sec, and 72°C for 3 min; followed by 25 cycles of 94°C for 30 sec, 68°C for 30 sec and 72°C for 3 min. Second round RACE PCR utilized 1 μl of first round RACE products as template amplified under the same conditions using nested, gene-specific primers. RACE PCR products were purified and ligated into pGemT, and eight to ten clones were selected for sequencing.

2.25 Construction of recombinant adenoviruses

Three synthetic variants of UL139 with FLAG-tags inserted internally or at the C terminus were provided commercially by GenScript and supplied as stab cultures containing the gene in the plasmid pUC57 (six tagged variants in total). The cultures were inoculated onto L-agar (containing 100 μg/ml ampicillin) plates and incubated at 37°C overnight. A single colony of each was inoculated into 2YT (containing 100 μg/ml ampicillin) and incubated at 37°C overnight with shaking (180 rpm). Plasmid DNA was purified using a Qiaprep miniprep kit according to the manufacturer's instructions (Qiagen) and quantified using a spectrophotometer.

The tagged gene [containing regions of homology with the adenovirus vector (pAL942)] was excised by digesting plasmid DNA with XbaI and BamHI. The excised band was purified using a Geneclean turbo spin kit according to the manufacturer's instructions (Q-biogene) and quantified using a spectrophotometer. The DNA was electroporated into freshly prepared competent SW102 cells and the cells were recovered in 5 ml of LB at 32°C with shaking for 4 h (180 rpm). SW102 cells contain pAL942, the adenovirus BAC vector with which the tagged UL139 variants were to recombine, and the lambda red proteins (including Redα, a double strand specific 5′ to 3′ exonuclease, Redβ which mediates strand exchange and annealing and the λGam protein which protects the ends of its linear genome from degradation) that mediate recombination.

A 50 μl volume of each sample (and a number of tenfold dilutions, 1:10, 1:100 and 1:1000) were plated on L-agar [without NaCl, with sucrose (6% w/v), chloramphenicol (40 μg/ml), IPTG (10 μl) and X-gal (20 μl)] plates and incubated at 32°C for 36 h. pAL942 encodes β-galactosidase, an enzyme that converts X-gal into a blue product. Recombination between the tagged UL139 variants and pAL942 results in interruption of the β-galactosidase gene. White bacterial colonies indicated that the β-galactosidase gene had been interrupted and that recombination had occurred, whereas blue colonies indicated an intact β-galactosidase gene. Independent white colonies were selected, inoculated into LB (containing appropriate antibiotic) and tested for the presence of UL139-tagged inserts by PCR using PMV100 forward and reverse primers (Table 2.2).

2.26 Western blotting

2.26.1 Preparation of proteins

HFFF-2 cells (10^6) were seeded into a 25 cm^3 tissue culture flask and incubated at 37°C overnight. The medium was removed and replaced with 1.5 ml of fresh medium. The cells were infected or mock-infected at an m.o.i. of 100 p.f.u./cell and incubated at 37°C with rocking for 1 h. The medium was removed and the cells were washed with 2 ml of fresh media. Fresh medium (5 ml) was then added to each flask and incubated at 37°C for 72 h.

The medium was removed and the cells were washed with 5 ml of ice-cold PBS complete. Ice-cold PBS complete (4 ml) was added alongside pre-chilled glass beads and the flask was shaken to detach the cells. The flask contents were transferred to a pre-chilled Falcon tube through a sieve. The glass beads and flask were washed with PBS complete and all washings were transferred to the Falcon tube, which was then centrifuged at 835 x g for 10 min at 4°C. The supernatant was discarded and the cell pellet was resuspended in 1 ml of ice-cold PBS complete. This was then transferred to a pre-chilled microcentrifuge tube and centrifuged at 6,500 x g for 1 min at 4°C. The supernatant was discarded and the cell pellet was resuspended in 130 μl of ice-cold PBS complete. A 30 μl aliquot was mixed with 30 μl of 1 X loading buffer and boiled for 10 min to denature proteins. The remainder of the sample was stored at -20°C.

2.26.2 SDS-PAGE

The Bio-Rad mini-protean II cell apparatus was used for the preparation of SDS-PAGE gels. The running gel (15% polyacrylamide) was prepared as follows:

37.5% (w/v) acrylamide	6.25 ml
RGB	3 ml
TEMED	10 μl
APS 10% (w/v)	100 μl
Sterile distilled water	2.25 ml

The running gel (7.5 ml) was poured and water was added to ensure a level surface at the top of the gel. The gel was allowed to set for about 30 min. The water was drained off and excess water dried using 3MM paper. The stacking gel was prepared as follows:

37.5% acrylamide	0.4 ml
SGB	0.6 ml
TEMED	3 µl
APS 10%	20 µl
Sterile distilled water	1.4 ml

The comb was inserted and the stacking gel was poured, ensuring that no bubbles remained. The gel was allowed to set for about 30 min. A 10 µl aliquot of each protein sample and 3 µl of rainbow protein marker (Amersham Biosciences) was loaded onto the gel, which was then electrophoresed at 100 V until the dye front had reached the bottom of the gel.

2.26.3 Transfer of proteins to membrane

The Bio-Rad mini trans-blot apparatus was used to transfer proteins from the gel to an ECL nitrocellulose membrane. The gel was removed from the casting plates and placed in transfer buffer. 3MM paper and membrane were cut to size. The membrane was placed in methanol for 15 sec, water for 5 min and transfer buffer for 5 min. 3MM paper and foam pads were pre-soaked in transfer buffer. The apparatus was assembled for transfer as follows: (black side of cassette down), sponge, filter paper, gel, membrane, filter paper, sponge pad. Proteins were transferred at 100V for 120 min, using an ice pack to cool.

2.26.4 Antibody detection of proteins

The membrane was washed in 1 X PBST for 5 min at RT with shaking. The membrane was then blocked in 20 ml of blocking solution for 2 h at RT with shaking. The blocking solution was discarded and the membrane was sealed in a bag with 10 ml of blocking solution containing primary antibody at an appropriate dilution. The bag was incubated at 4°C with shaking overnight. The blot was removed from the bag and washed (three times) in 20 ml of PBST. The secondary antibody diluted in blocking solution was then added to each blot, sealed in another bag and incubated at RT with shaking for 1 h. The blot was washed (three times) in 20 ml of PBST. ECLI and II (GE healthcare) were mixed (ratio 40:1) and 1 ml was added (per blot), spread evenly and incubated for 40 sec. Excess ECL was drained and the blot was sealed in a bag and exposed to photographic film for varying lengths of time to obtain a range of images.

3 Genotyping of UL146 and UL139

3.1 Introduction

Calculation of nucleotide divergence between nine HCMV strains, using an alignment of the sequences at the right end of U_L (Chapter 1, Figure 1.5), revealed that UL146 and UL139 are the two most hypervariable genes in this region (Dolan *et al.,* 2004). In order to ascertain whether this was true for a larger number of strains, nucleotide divergence was calculated for 27 HCMV strains (Figure 3.1). The sequences were aligned using ClustalW and the consensus sequence was extracted, with any nucleotide position that differed between one or more strains and the others, plus any gaps, counted as divergent and represented by hyphens. The distribution of hyphens was counted using the GCG programme Window and the results were plotted using the GCG programme StatPlot (Figure 3.1). As observed in the earlier analysis, UL146 and UL139 show the greatest level of nucleotide divergence in this region. Indeed, for these genes nucleotide divergence levels approach 100% because the alignment process breaks down. Other genes in this region, such as UL144, also show substantial levels of nucleotide divergence, although these are less marked than in UL146 and UL139.

The literature documents 14 UL146 genotypes (in 26 isolates; Dolan *et al.,* 2004) and six UL139 genotypes (in 26 isolates; Qi *et al.,* 2006). The present study investigated circulating genotypes of both genes in a much larger panel of clinical isolates from diverse geographical and clinical settings. The sequences were analysed to investigate genotypic frequencies, geographical distribution of genotypes, and possible modes of evolution. The predicted protein sequences of the 14 UL146 genotypes were used to build homology models in order to test to what extent sequence differences are likely to result in structural differences.

3.2 UL146 and UL139 sequences

A collection of DNA extracts from 184 virus samples was established, consisting of 179 anonymised clinical samples and five commonly used laboratory strains. Details of the samples (171) successfully amplified by PCR for at least one of the genes are summarized in Table 3.1 and include, where known, the clinical

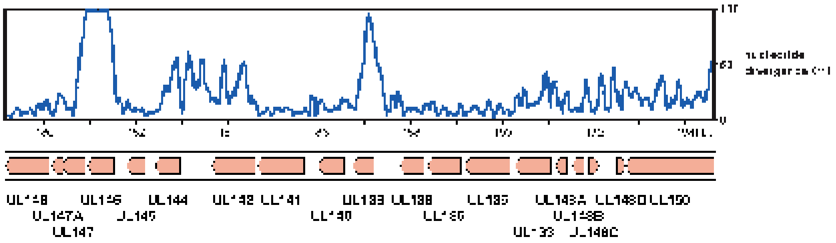


Figure 3.1 Nucleotide sequence divergence at the right end of U_L in 27 HCMV strains

Nucleotide divergence between 27 HCMV strains, the nine strains in Figure 1.5, plus published data for strains AL, NT, 5234, 711, (Dolan *et al.,* 2004), TR and PH (Murphy *et al.,* 2003), plus unpublished data for strains VR3216B, 4119, VR1814, U3, 309, 66, Cincy, JP, U4, AF1, U11 and U8 (courtesy of A. Davison) was calculated using an alignment of the sequences of the right end of U_L. If all strains were not identical, a nucleotide position was counted as divergent. as were gaps in the alignment. The plot shows nucleotide divergence in a 100 nucleotide window shifted by increments of three nucleotides. The protein-coding regions in this region of the genome are shown below the plot, with a scale based on their position in strain Merlin (Chapter 1, Figure 1.5). Figure provided by A. Davison.

Table 3.1: Information on HCMV Samples Used and Genotypes Determined

Strain[a]	*Source*[b]	*Age/sex*[c]	*Details*[d]	*UL146 genotype*[e]	*UL139 genotype*[e]
A1	U	Adult/M	I/CL	G2	ND
A2	U	7/M	AA	G9	ND
A3	U	46/M	R	G4 (-)	G8 (-)
A4	U	9/M	B	G9 (G13)	G7 (-)
A5	U	1/F	B	G4 (-)	G2, G7
A6	U	15 d/F	J/C	G13 (-)	G1, G4, G7 (-)
A7	U	1/M	SCID/C/B	G13	G1
A8	U	8 m/M	C	ND	G2
A9	U	2 m/F	I	G13	G2
A10	U	1/M	SCID/C/B	G13	G4
A11	U	Infant	C	ND	G4, G6
A12	U	Infant	?	G7 (-)	G2 (-)
A13	U	Infant	?	ND	G4
A14	U	Child	B	G5 (-)	G7 (-)
A15	U	Infant	C	ND	G7
A16	A	Adult/F	C/C+	G2	G2
A17	A	Adult/F	C/C+	G9	ND
A18	TS	Infant	C/I	G13	G4
C1	U	Adult	KT	ND (G12)	G5 (G4)
C2	U	Infant	C	G2 (G9)	G6 (G3, G5)
C3	U	Infant	C	G6, G9 (G7)	G3 (G1, G4)
C4	U	Infant	C	ND (G7)	G4 (-)
C5	U	Infant	C	G7 (-)	ND
C6	U	Infant	C	G7 (-)	ND (G4)
C7	U	Infant	C	G7 (-)	G3 (G4)
C8	U	Infant	C	G1 (-)	ND (G2, G3, G8)
C9	U	Infant	C	ND (-)	G5 (-)
C10	U	Infant	C	G1 (G7)	G1 (G8)
D1	BL/P	?	?	G5, G9	ND
D2	BL/P	?	?	G7 (G3, G12, G13)	ND (G1, G4)
D4	BL/P	?	?	G12	G6
D5	BL/P	?	?	G12 (G4, G7)	ND (G2, G3, G7)
D6	BL/P	?	?	G8	G5
D7 *(TB40/E)*	TS/P	?	B	G8	G4
E1	?	Infant	?	G7	ND
E2	?	R	?	ND	ND
E3	?	T	?	G7	ND
E4	?	Infant	?	G13 (-)	G7 (G1)
E5	?	3	?	G9 (G7, G12)	ND (G4)
E6	U	Infant/M	T	G2	G1, G2
E7	U	Adult/M	LT	ND (G13)	G4, G6 (-)
E8	Ti	Adult	AIDS	G10	G2
E9	U	Infant	C	ND	G4
E10 *(AL)*	LT	Adult	H	G10 (-)	G5 (G4)
E11 *(NT)*	T	Adult	H	G4	G3
E12 *(W)*	LT	Adult	H	G13 (G7)	G1, G2, G7 (-)
E13	U	?	LT	G9	ND
G1	U	Infant	C	G4 (-)	G4 (-)
G2	U	Infant	C	G13 (-)	G2, G3 (-)
G3	U	Infant	C	ND	G2, G7
G4	U	Infant	C	G13 (-)	G4 (G7)
G5	U	Infant	C	G13	G8
G6	U	Infant	PN	G14	G1
G7	U	Infant	PN	ND	G2
G8	U	Infant	PN	G5 (G12, G13)	G4 (G1, G2)
G9	U	Infant	PN	G2	ND
G10	U	Infant	PN	G5	G2
G11	U	Infant	PN	G9	ND

G12	U	Infant	PN	G13	G4
G13	U	Infant	PN	ND	G7
G14	U	Infant	PN	G7 (-)	G1 (-)
G15	U	Infant	PN	G2	ND
G16	U	Infant	PN	ND	ND
G17	U	Infant	PN	G8	G1, G5, G8
G18	U	Infant	PN	G3 (G7, G8, G12)	G1 (-)
H1	U	Infant	C	G11	G2
H2	U	Infant	C	G9	ND
H3	U	Infant	C	G9	G4, G7
H4	U	Infant	C	G11	G2
H5	U	Adult/F	C+	G13	ND
H6	U	66/M	CH	G10	G4
H7	U	55/M	C+	G1	G4
H8	U	Adult/F	C+	ND	ND
H9	U	Infant	C	G7	G2
H10	U	Infant	C	G7	G4
H12	U	Infant	C	G12	G5
H13	U	Adult/F	C+	ND	ND
H14	U	Infant	C	G8	G7
H15	U	Adult/F	C+	ND	G1
H16	U	Infant	C	G11	G4
H17	U	Infant	C	ND	G1
H18	U	Infant	C	G13	G4
H19	U	Adult/F	C+	ND	G4
H20	U	Infant	C	ND	ND
H22	U	Infant	C	ND	G2
H26	U	10 w/F	C+	ND	G1
H27	U	Adult/F	C+	ND	ND
H28	U	Adult/F	C+	G2	ND
H29	U	14 m/M	C+	G10	G2
I1	U	5 m/M	C	G9	ND
I2	U	2 d/M	C	G8 (-)	ND (-)
I3	C	Adult/F	P	G7, G13	G4
I4	A	Adult/F	?	ND	G8
I5	U	4 m/M	?	G12 (-)	G4 (G5)
I6	B	Adult/M	HT	G7	G2
I7	C	Adult/F	P	G13	G4
N1	U	Adult	R	G1	G4
N2	U	Adult	R	G12	G3
N3	U	Infant	C	G7	G2
N4	U	Infant	C	G7	G6
N5	U	Infant	C	ND	G2
N6	U	Infant	C	ND	G2
S1	U	1/M	C+	G1, G7	G1
S2	U	45/F	C+	G7 (G13)	G2 (G4)
S3	U	10/M	B	G7, G12	G2
S4	P	35/M	B	G9, G12, G13 (G2, G7)	G2, G5 (G1)
S5	P	43/M	B	G13	ND
S7	BL	56/M	AP	G7 (-)	G6 (G5)
S8	BL	45/M	RF	G14	ND
S9	SR	58/M	R	ND	ND
S10	P	66/M	H	ND	G4
S12	U	2/M	W	G7 (G2, G9)	G2 (G6)
S13	TS	0/F	P	G2 (G4)	G2 (G6)
S14	U	29/F	P	G7	ND
S15	P	58/F	?	ND	ND
S16	TS	32/F	?	ND	ND
S17	P	38/F	H	ND	ND
S18	SR	46/F	A	ND	ND

S20	SR	31/F	L	ND	G5
S21	BL	81/M	?	G12	ND
S22	F	75/F	PC	G10	ND
S23	P	49/M	R	G10 (G12)	G5 (G4)
S24	P	68/F	R	ND	G5
S25	BP	19/M	R	G5	ND
S26	BP	48/F	CCO	G1, G10, G13	G5
S27	U	14 w/M	C	G12	G5
S28	P	36/M	N/IM	G3	G1
S30	E	56/F	CO	G4	ND
S31	BL	58/M	L	G13	G5
S32	U	19/F	?	ND	ND
S33	SP	40/M	?	G13 (G9, G12)	G4 (G1, G5)
S34	P	44/F	?	G12	G2
S35	TS	68/M	?	G7 (-)	ND (G3)
S36	TS	9/M	AL	ND	ND
S37	P	65/M	R	ND	G3
S39	U	?	KT	G13	G2
S40	U	Adult/M	KT	ND (G12)	G1 (-)
S41	B	Adult/M	KT	G13	G1
S42	U	Infant/M	C	G10	G2
S43	U	Adult	KT	G13	G1, G4
S44	U	Adult	KT	G2	G6, G7
S45	U	?	KT	G1 (-)	G6 (G3)
S46	U	Adult	KT	ND	G2
U1 *(Cincy)*	BAC	Adult	AIDS	G9	G2
U2 (Ad*var*UC)	AD	Child	?	G9	G7
U3 *(Davis)*	BP	Infant	C	G5 (-)	G4 (-)
U4 (Toledo)	U	Infant	C	G1 (-)	G4 (-)
U5 (Towne)	U	Infant	C	G7	G5
W2 *(711)*	U	Infant	C	G1	G5
W3 *(4119)*	U	Infant	C	G8	G2
W4 *(5234)*	A	Infant	C	G1	G1
W5 *(6397)*	U	Infant	C	G12	G4
W6 *(3301)*	U	Infant	C	G9	G6
W7	A	Infant	C	G9	G6
W8 *(3157)*	U	Infant	C	G7	G6
W9 *(Merlin)*	U	Infant	C	G2 (-)	G1 (-)
Z1	S	28/F	H/C+	ND	G2, G5
Z2	S	36/F	C+	G1, G13 (G12)	G5 (G1, G4)
Z3	S	22/F	C+	G9, G13 (-)	G1 (-)
Z4	S	20/F	C+	G9 (G5)	G2, G5 (-)
Z5	S	18/F	C+	G3, G14	G2
Z6	S	20/F	C+	G13 (G1, G7)	ND (G4, G5)
Z7	S	29/F	C+	G1, G3, G9 (G7)	G4 (G5)
Z8	S	22/F	C+	ND	G5
Z9	S	21/F	C+	G9	ND
Z10	S	26/F	H/C+	ND	G5
Z11	S	26/F	H/C+	G1 (G12)	G1 (G4, G5)
Z12	S	21/F	H/C+	G9	ND
Z13	S	30/F	H/C+	G13	ND
Z14	S	30/F	C+	G3, G7, G13 (G12)	G5 (G2)
Z15	S	23/F	C+	G13	ND

[a] A, Australia; C, Hong Kong; D, Germany; E, England; G, The Gambia; H, Hungary; I, Italy; N, The Netherlands; S, Scotland; U, USA; W, Wales; Z, South Africa. The original strain designations of sequences listed in a previous study (Dolan *et al.,* 2004) are given in parentheses.

[b] A, amniotic fluid; AD, adenoids; B, blood; BAC, bacterial artificial chromosome; BL, bronchoalveolar lavage; BP, biopsy; C, cervical swab; E, endotracheal swab; F, faeces; LT, lung tissue; P, plasma; /P, passaged in cell culture; S, saliva; SP, sputum; SR, serum; T, thyroid; Ti, tissue (unspecified); TS, throat swab; U, urine; ?, unknown.

[c] Ages are in years unless specified in days (d), weeks (w) or months (m). Sexes: F, female; M, male; ?, unknown.

[d] AIDS, acquired immunodeficiency syndrome; A, acute myeloid leukaemia; AA, aplastic anaemia; AL, acute lymphoblastic leukaemia; AP, aplastic anaemia with pneumonia; B, bone marrow transplant; C, congenital; C+, HCMV positive by PCR or IgM in serum; CCO, Crohn's disease with HCMV colitis; CH, chronic haemodialysis patient; CL, chronic lung disease; CO, colitis; H, HIV positive; HT, heart transplant; IM, immunosuppressed; J, jaundice; KT, kidney transplant; L, lower respiratory tract infection; LT, liver transplant; N, nephrotic syndrome; P, maternal HCMV positive (pregnancy); PC, pan-proctocolectomy; PN, postnatal; R, renal transplant; RF, respiratory failure; SCID, severe combined immunodeficiency; T, thrombocytopenia; W, Wilms' tumour post-nephrectomy.

[e] Genotypes are denoted G1-G14 for UL146 and G1-G8 for UL139. Multiple genotypes are separated by commas. ND, not determined. Additional genotypes identified in subsequent experiments are in parentheses; -, no additional genotypes identified.

source of the original sample, the age and sex of the patient and clinical details regarding the medical condition of the patient. Details of collaborators who collected and provided the samples and extracted the DNA are shown in Table 2.1 (Chapter 2).

3.3 UL146 genotypes

A total of 184 HCMV DNA samples were tested for the presence of UL146. UL146 was amplified successfully as a PCR product from 159 samples, and sequences were determined successfully from 134 PCR products (Table 3.1). UL146 (and UL139, see Section 3.4) were not amplified from 13 samples, which are excluded from Table 3.1. Some samples contained more than one sequence, and in total the 134 samples yielded 182 UL146 sequences. The phylogenetic and diversity analyses involved a total of 350 UL146 sequences, which included the 182 sequences derived from the present study (Table 3.1) and all UL146 sequences reported by others in the literature or deposited in GenBank (Table 3.2).

The UL146 coding sequences range in length from 342-378 bp (114-126 codons, including the stop codon).

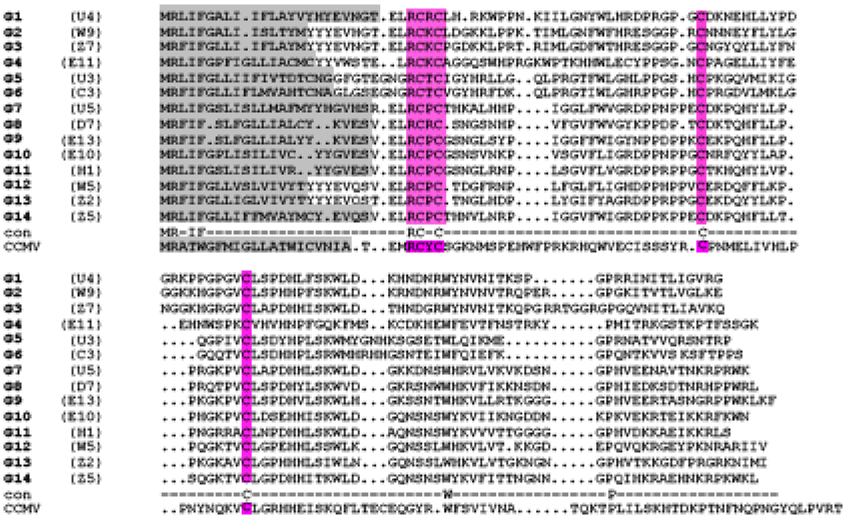
Table 3.2: Additional UL146 sequences Used for Phylogenetic and Diversity Analyses

Accession numbers	*Reference*
AY681088 - AY681116	Arav-Boger *et al.* (2005, 2006)
AY446877-AY446893	Dolan *et al.* (2004)
DQ115708-DQ115756	Lurain *et al.* (2006)
AY582483-AY582530	Stanton *et al.* (2005)
DQ229942-DQ229946, DQ229948	Ruan *et al.* (unpublished)
DQ180366	Zhou *et al.* (unpublished)
AY788113-AY788136	He *et al.* (2006)

Amino acid sequence alignments and phylogenetic trees produced by the neighbour-joining method were employed to group the sequences into genotypes. All of the sequences fell into the 14 UL146 genotypes (G1-G14) defined previously (Dolan *et al.,* 2004). An aa sequence alignment containing a representative of each genotype is shown in Figure 3.2. The protein encoded by each genotype contains a putative signal peptide sequence (highlighted in grey).

UL146 is hypervariable throughout its length and only a few aas are completely conserved in the mature protein: three residues in the RCXC motif, two other C residues, and single W and P residues. All genotypes contain the characteristic RCXC motif, which is present as ELRCXC in 12 genotypes and as NGRCXC in two genotypes (G5 and G6). Table 3.3 displays a laboratory strain or previously published strain (Dolan *et al.,* 2004) representing each of the fourteen genotypes.

An unrooted phylogenetic tree showing the relationship between the HCMV UL146 genotypes and CCMV UL146 is shown in Figure 3.3. Bootstrap values under 70 indicate regions of unresolved branching order. The genotypes cluster into four groups. Group A contains the most members but the order of genotypes within this group is not clear. However, G10 and G11 are more closely related to each other than to other members of the group, as are G12 and G13 and, similarly, G8 and G9. Group B contains three members, G1, G2 and G3. Group C contains two members, G5 and G6. In group D, G4 appears to be more closely related to the CCMV UL146 sequence than to any of the other HCMV genotypes.



```
G1    (U4)   MRLIFGALI.IFLAYVYHYEVNGT.ELRCRCLH.RKWPPN.KIILGNYWLHRDPRGP.GCDKNEHLLYPD
G2    (W9)   MRLIFGALI.ISLTYMYYYEVNGT.ELRCKCLDGKKLPPK.TIMLGNFWFHRESGGP.RCNNNEYFLYLG
G3    (Z7)   MRFIFGLLI.IFLAYMYYYEVNGT.ELRCKCPGDKKLPRT.RIMLGDFWTHRESGGP.GCNGYQYLLYFN
G4   (E11)   MRLIFGPFIGLLIACMCYYWSTE..LRCKCAGGQSWHPRGKWPTKHHWLECYPPSG.HCPAGELLIYFE
G5    (U3)   MRLIFGLLIIFIVTDTCNGGFGTEGNGRCTCIGYHRLLG..QLPRGTFWLGHLPPGS.HCPKGQVMIKIG
G6    (C3)   MRFIFGLLIFLMVAHTCNAGLGSEGNGRCTCVGYHRFDK..QLPRGTIWLGHRPPGP.HCPRGDVLMKLG
G7    (U5)   MRLIFGSLISLLMAFMYYHGVHSR.ELRCPCTHKALHHP....IGGLFWVGRDPPNPPECDKPQHYLLP.
G8    (D7)   MRFIF.SLFGLLIALCY..KVESV.ELRCRC.SNGSNHP....VFGVFWVGYKPPDP.TCDKTQHFLLP.
G9   (E13)   MRFIF.SLFGLLIALYY..KVESV.ELRCPCGSNGLSYP....IGGFFWIGYNPPDPPKCEKPQHFLLP.
G10  (E10)   MRLIFGPLISILIVC..YYGVESV.ELRCPCGSNSVNKP....VSGVFLIGRDPPNPPGCNRFQYYLAP.
G11   (H1)   MRLIFGSLISILIVR..YYGVESV.ELRCPCGSNGLRNP....LSGVFLVGRDPPRPPGCTKHQHYLVP.
G12   (M5)   MRFIFGLLVSLVIVYTYYYEVQSV.ELRCPC.TDGFRNP....LFGLFLIGHDPPHPPVCERDQFFLKP.
G13   (Z2)   MRFIFGLLIGLVIVYTYYYEVQST.ELRCPC.TNGLHDP....LYGIFYAGRDPPRPPGCEKDQYYLKP.
G14   (Z5)   MRLIFGLLIFFMVAYMCY.EVQSV.ELRCPCTHNVLNRP....IGGVFWIGRDPPKPPECDKPQHFLLT.
con          MR-IF----------------------RC-C----------------------------C----------
CCMV         MRATWGFMIGLLATMICVNIA.T..EHRCYCSGKNMSPEHWFPRKRHQWVECISSSYR.CPNMELIVHLP

G1    (U4)   GRKPPGPGVCLSPDHLFSKWLD...KHNDNRWYNVNITKSP.......GPRRINITLIGVRG
G2    (W9)   GGKKHGPGVCLSPHHPFSKWLD...KRNDNRWYNVNVTRQPER.....GPGKITVTLVGLKE
G3    (Z7)   NGGKHGRGVCLAPDHHISKWLD...THNDGRWYNVNITKQPGRRTGGRGPGQVNITLIAVKQ
G4   (E11)   ..EHNWSPKCVHVHNPFGQKFMS..KCDKHEWFEVTFNSTRKY......PHITRKGSTKPTFSSGK
G5    (U3)   ....QGPIVCLSDYHPLSKWNYGNHKSGSETMLQIKME..........GPRNATVVQRSNTRP
G6    (C3)   ....GQQTVCLSDHHPLSRWNHRHHGSNTEIMFQIEFK..........GPQNTKVVS KSFTPPS
G7    (U5)   ...PRGKPVCLAPDHHLSKWLD...GKKDNSWHRVLVKVKDSN.....GPHVEENAVTNKRPRWK
G8    (D7)   ...PRQTPVCLSPDHYLSKWVD...GKRSNWWHKVFIKKNSDN.....GPHIEDKSDTNRHPPWRL
G9   (E13)   ...PKGKPVCLSPDHVLSKWLH...GKSSNTWHKVLLRTKGGG.....GPHVEERTASNGRPPWKLKF
G10  (E10)   ...PHGKPVCLDSEHHISKWLD...GQNSNSWYKVIIKNGDDN.....KPKVEKRTEIKKRFKWN
G11   (H1)   ...PNGRRACLNPDHHLSKWLD...AQNSNSWYKVVVTTGGGG.....GPHVDKKAEIKKRLS
G12   (M5)   ...PQGKTVCLGPEHHLSSWLK...GQNSSLWHKVLVT.KKGD.....EPQVQKRGEYPKNRARIIV
G13   (Z2)   ...PKGKAVCLGPHHHLSIWLN...GQNSSLWHKVLVTGKNGN.....GPHVTKKGDFPRGRKNIMI
G14   (Z5)   ...SQGKTVCLGPDHHITKWLD...GQNSNSWYKVFITTNGNN.....GPQIHKRAEHNKRPKWKL
con          ---------C---------------------W-----------------P-----------------
CCMV         ..PNYNQKVCLGRHHEISKQFLTECEQGYR.WFSVIVNA......TQKTPLILSKHTDKPTNFNQPNGYQLPVRT
```

Figure 3.2 Amino acid sequence alignment of UL146 genotypes

An alignment of amino acid sequences representing the 14 genotypes (G1-G14), with representative strains (Table 3.1) indicated in parentheses. Completely conserved residues are shown in the consensus row (con) and non-conserved residues are represented by hyphens. CCMV UL146 is shown below the consensus. Predicted signal sequences are highlighted in grey. The RCXC motif and two additionally conserved cysteine residues are highlighted in pink.

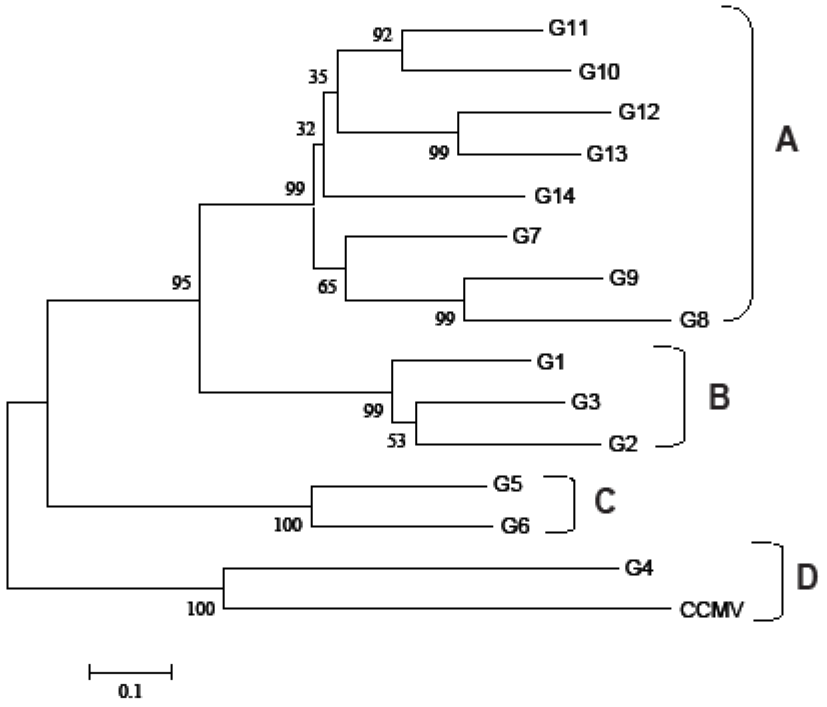


Figure 3.3 Phylogenetic tree of UL146 genotypes

An unrooted neighbour-joining phylogenetic tree of representatives of the 14 UL146 genotypes (G1-G14) and CCMV UL146 was produced in Mega 4.0, using the amino acid sequence alignment shown in Figure 3.2. Genotypes cluster into four groups, A, B, C and D.The scale bar indicates divergence as substitutions per amino acid site. Bootstrap values are out of 100.

Table 3.3: Laboratory strain corresponding to UL146 Genotype

UL146 Genotype	*Laboratory strain*
G1	Toledo
G2	Merlin
G3	KSG
G4	NT
G5	Davis
G6	ML1
G7	Towne
G8	TB40E
G9	FS
G10	Al
G11	[F]
G12	6397
G13	KM
G14	RK

Genotypic frequencies were calculated from the total number of UL146 sequences available (from the present study and in Genbank, a total of 350) (Figure 3.4). UL146 G7 and G9 were each detected in 14% of sequences or more, and UL146 G12 and G13 in over 12%. In contrast, UL146 G6 and G14 were each detected in less than 2% of sequences. The remainder of the UL146 genotypes were detected at frequencies of 2-10%.

Amino acid sequence alignments were used to calculate the level of sequence identity between all strains within each individual genotype (using Swaap) (Figure 3.5). An aa sequence alignment (Figure 3.2) of a representative of each UL146 genotype was used to calculate identity among genotypes. The aa sequence identity among genotypes is low (range 18.5-68.2%, mean 36.9%), whereas within each genotype it is high (range 94.9-99.5%). Similarly, nucleotide sequence identity is low among genotypes (range 47.2-81.2%, mean 55.4%), whereas it is high within each genotype (range 97.2-99.8%). In contrast to the high level of identity within genotypes, identity within each of the four groups was low; therefore the four groups were not employed in further analyses.

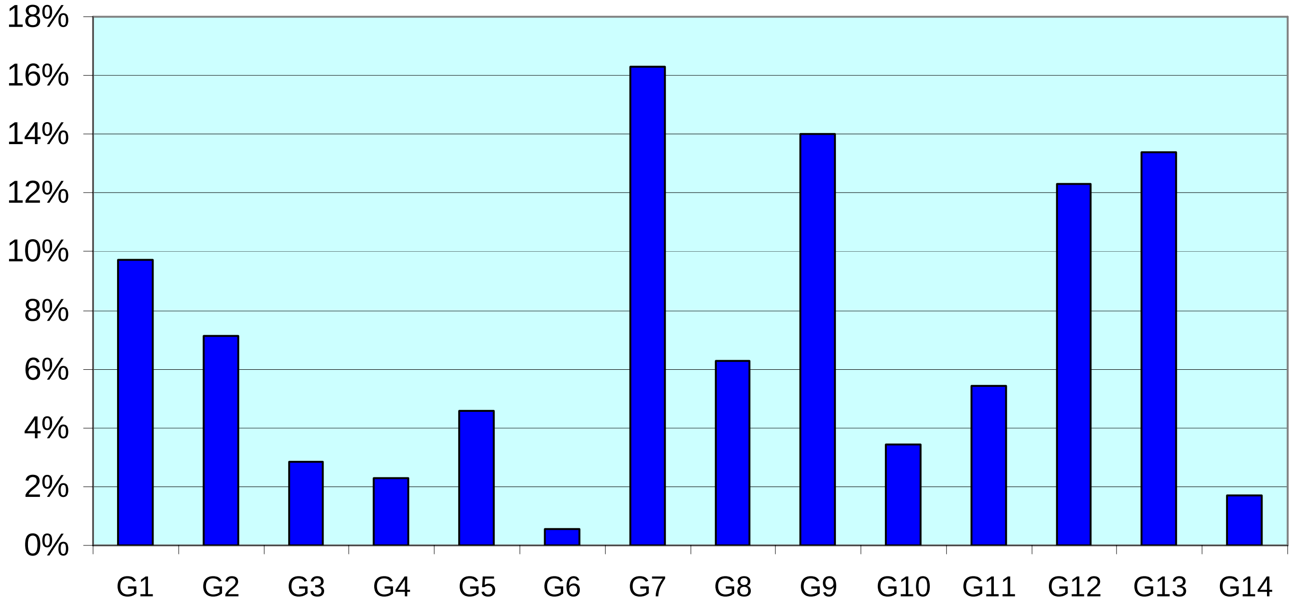


Figure 3.4 Frequencies of occurrence of UL146 genotypes
The frequencies were calculated using all available sequences from the present study and those available in GenBank (a total of 350).

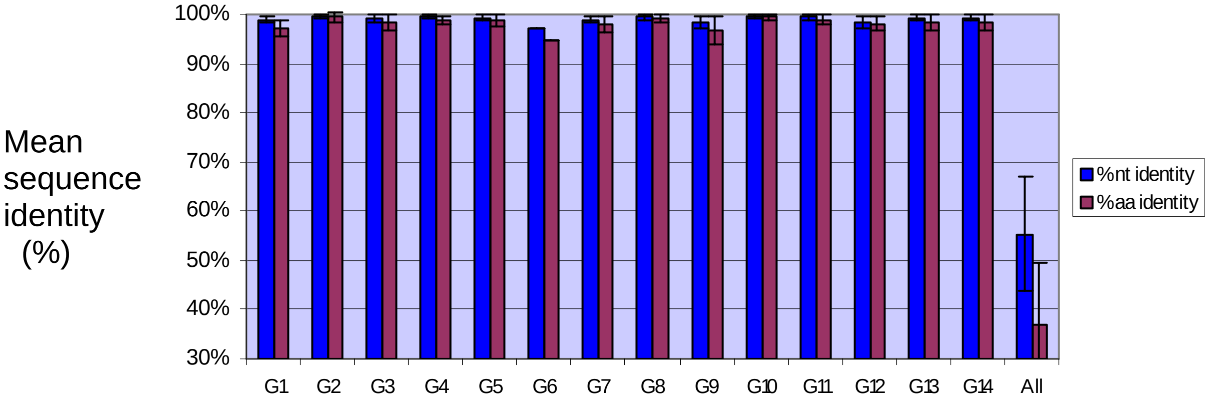


Figure 3.5 Sequence identity within and among UL146 genotypes

Nucleotide (nt) and amino acid (aa) sequence identities were calculated *within* each genotype (G1-G14) by pairwise alignment of all sequences in the relevant genotype, and *among* all genotypes (All) by pairwise alignment of a representative of each genotype (Figure 3.2). Mean and standard deviation values are shown.

```
N1      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
20M     MRLIFGALIIFLAYVYHYEVNGTELRCRCLHKKWPPNKIILGNYWLHRDPRGPGCDKNE
2J      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHKKWPPNKIILGNYWLHRDPRGPGCDKNE
W4      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
W2      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
A9      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGSYWLHRDPRGPGCDKNE
C3      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
CH11    MRLIFGALIIFLAYVYHYEVNGTELRCRCLHKKWPPNKIILGNYWLHRDPRGPGCDKNE
CH18    MRLIFGALIIFLAYVYHYEVNGTELRCRCLHKKWPPNKIILGNYWLHRDPRGPGCDKNE
CH19    MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
CH2     MRLIFGALIIFLAYVYHYEVNGTELRCRCLHKKWPPNKIILGNYWLHRDPRGPGCDKNE
CH20    MRLIFGALIIFLAYVYHYEVNGTELRCRCLHKKWPPNKIILGNYWLHRDPRGPGCDKNE
CH21    MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
CH3     MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
CH8     MRLIFGALIIFLAYVYHYEVNGTELRCRCLHKKWPPNKIILGNYWLHRDPRGPGCDKNE
H7      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
S1      MRLFSGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
S26     MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
NA      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
PT12    MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
PT16    MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
S11     MRLLFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
Z11     MRLLFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
Z2      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPSNKIILGNYWLHRDPRGPGCDKNE
Z6      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
Z7      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
SR      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
U4      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
TR      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
S45     MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
C8      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
C10     MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
UU      MRLIFGALIIFLAYVYHYEVNGTELRCRCLHRKWPPNKIILGNYWLHRDPRGPGCDKNE
con     MRL--GALIIFLAYVYHYEVNGTELRCRCLH-KWP-NKIILG-YWLHRDPRGPGCDKNE

N1      HLLYPDGRKPPGPGVCLSPDHLFSKWLDKYNDNRWYNVNITKSPGPRRINITLIGVRG
20M     HLLYPDGRKPPGPGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
2J      HLLYPDGRKPPGPGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
W4      HLLYPNGRKPP..GVCLSPDHLFSKWLDKHDDNRWYNVNITKSPGPRRINITLIGVGG
W2      HLLYLDGRKPPGPGVCLSPDHLFSKWLDKHNDDRWYNVNITKSPGPRRINITLIGVRG
A9      HLLYPDGRKPPGPGVCLSPDHLFSKWLDKHNDNRWYNVNITKSPGPRRINITLIGVRG
C3      HLLYPDGRKPPGPGVCLSPDHLFSKWLDKYNDNRWYNVNITKSPGPRRINITLIGVKG
CH11    HLLYPDGRKPPGPGVCLSPDHLFSKWLDKYNDNRWYNVNITKSPGPRRINITLIGVKG
CH18    HLLYPDGRKPPGPGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
CH19    HLLYPDGRKPPGSGVCLSPDHLFSKWLDKYNDNRWYNVNITKSPGPRRINITLIGVRG
CH2     HLLYPDGRKPPGPGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
CH20    HLLYPDGRKPPGPGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
CH21    HLLYPNGKKPP..GVCLSPDHLFSKWLDKHDDNRWYNVNITKSPGPRRINITLIGVGG
CH3     HLLYPDGRKPPGPGVCLSPDHLFSKWLDKYNDNRWYNVNITKSPGPRRINITLIGVKG
CH8     HLLYPDGRKPPGPGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
H7      HLLYPDGRKPPGPGVCLSPDHLFSKWLDKHNDNRWYNVNITKSPGPRRINITLIGVRG
S1      HLLYPDGRKPPGHGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
S26     HLLYPDGRKPPGHGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
NA      HLLYPNGKKPP..GVCLSPDHLFSKWLDKHDDNRWYNVNITKSPGPRRINITLIGVGG
PT12    HLLYPDGRKPPGHGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
PT16    HLLYPDGRKPPGPGVCLSPDHLFSKWLDKYNDNRWYNVNITKSPGPRRINITLIGVRG
S11     HLLYPDGRKPPGHGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
Z11     HLLYPDGRKPPGHGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
Z2      HLLYPDGRKPPGHGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
Z6      HLLYPDGRKPPGPGVCLSPDHLFSKWLDKYNDNRWYKVNITKSPGPRRINITLIGVRG
Z7      HLLYPDGRKPPGHGVCLSPDHLFSKWLDKRNDNRWYNANMTKSPEPRRINITLIGVRG
SR      HLLYPDGRKPPGPGVCLSPDHLFSKWLDKYNDNRWYNVNITKSPGPRRINITLIGVRG
U4      HLLYPDGRKPPGPGVCLSPDHLFSKWLDKHNDNRWYNVNITKSPGPRRINITLIGVRG
TR      HLLYPDGRKPPGHGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
S45     HLLYPDGRKPPGHGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
C8      HLLYPDGRKPPGHGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
C10     HLLYPDGRKPPGHGVCLSPDHLFSKWLDKRNDNRWYNVNITKSPEPRRINITLIGVRG
UU      HLLYPDGRKPPGPGVCLSPDHLFSKWLDKYNDNRWYNVNITKSPGPRRINITLIGVRG
con     HLLY--G-KPP--GVCLSPDHLFSKWLDK--D-RWY--N-TKSP-PRRINITLIGV-G
```

Figure 3.1: Amino acid sequence alignment of UL146 G1 sequences

Figure 3.6 Amino acid sequence alignment of UL146 G1 sequences

All G1 sequences from the present study are included, plus those available in GenBank (in italics). Completely conserved residues are shown in the consensus row (con) and non-conserved residues are indicated by hyphens. Dots indicate gaps in the alignment. Mismatched residues are highlighted in yellow. The predicted signal peptide sequences are highlighted in grey, and the conserved RCXC motif and two cysteine residues are highlighted in pink.

```
A9      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
711     ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
U4      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
H7      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
Z6      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCACCTAATAAAATTA
N1      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCACCTAATAAAATTA
UU      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
PT16    ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
SR      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
CH19    ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
C3      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
CH3     ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
CH11    ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAAAAAATGGCCGCCTAATAAAATTA
S11     ATGCGATTACTTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
Z11     ATGCGATTACTTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
S1      ATGCGATTATTTTCTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
Z2      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGTCTAATAAAATTA
TR      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
S26     ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
S45     ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
C8      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
C10     ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
PT12    ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
Z7      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
2J      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAAAAAATGGCCGCCTAATAAAATTA
20M     ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAAAAAATGGCCGCCTAATAAAATTA
CH2     ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAAAAAATGGCCGCCTAATAAAATTA
CH20    ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAAAAAATGGCCGCCTAATAAAATTA
CH18    ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAAAAAATGGCCGCCTAATAAAATTA
CH8     ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAAAAAATGGCCGCCTAATAAAATTA
CH21    ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
NA      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
W4      ATGCGATTAATTTTTGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATAGAAAATGGCCGCCTAATAAAATTA
con     ATGCGATTA-TTT-TGGTGCGTTGATTATTTTTTTAGCATATGTGTATCATTATGAGGTGAATGGAACAGAATTACGCTGCAGATGTCTTCATA-AAAATGGCC--CTAATAAAATTA

A9      TATTGGGTAGTTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
711     TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCTAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCTGA
U4      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
H7      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
Z6      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
N1      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
UU      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
PT16    TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
SR      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
CH19    TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGATCTGGAGTATGTTTATCGCCCGA
C3      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
CH3     TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
CH11    TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
S11     TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACATGGAGTATGTTTATCGCCCGA
Z11     TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACATGGAGTATGTTTATCGCCCGA
S1      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACATGGAGTATGTTTATCGCCCGA
Z2      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACATGGAGTATGTTTATCGCCCGA
TR      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAGCCGCCTGGACATGGAGTATGTTTATCGCCCGA
S26     TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACATGGAGTATGTTTATCGCCCGA
S45     TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACATGGAGTATGTTTATCGCCCGA
C8      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACATGGAGTATGTTTATCGCCCGA
C10     TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACATGGAGTATGTTTATCGCCCGA
PT12    TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACATGGAGTATGTTTATCGCCCGA
Z7      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACATGGAGTATGTTTATCGCCCGA
2J      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
20M     TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
CH2     TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
CH20    TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
CH18    TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
CH8     TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAGACGGAAGGAAACCGCCTGGACCTGGAGTATGTTTATCGCCCGA
CH21    TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAAACGGAAAAAAACCGCCTGGA......GTATGTTTATCGCCCGA
NA      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAAACGGAAAGAAACCGCCTGGA......GTATGTTTATCGCCCGA
W4      TATTGGGTAATTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATCCAAACGGAAGGAAACCGCCTGGA......GTATGTTTATCGCCCGA
con     TATTGGGTA-TTATTGGCTTCATCGCGATCCCAGAGGGCCCGGATGCGATAAAAATGAACATTTATTGTATC-A-ACGGAA--AA-CCGCCTGGA------GTATGTTTATCGCC-GA
```

Figure 3.7 Nucleotide sequence alignment of UL146 G1 sequences
(continued overleaf)

```
A9        TCACCTCTTCTCAAAATGGTTAGACAAACACAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
711       TCACCTCTTCTCAAAATGGTTAGACAAACACAACGATGATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
U4        TCACCTCTTCTCAAAATGGTTAGACAAACACAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
H7        TCACCTCTTCTCAAAATGGTTAGACAAACACAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
Z6        TCACCTCTTCTCAAAATGGTTAGACAAATACAACGATAATAGGTGGTATAAGGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
N1        TCACCTCTTCTCAAAATGGTTAGACAAATACAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
UU        TCACCTCTTCTCAAAATGGTTAGACAAATACAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACTTTGATAGGTGTTAGAGGA
PT16      TCACCTCTTCTCAAAATGGTTAGACAAATACAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
SR        TCACCTCTTCTCAAAATGGTTAGACAAATACAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
CH19      TCACCTCTTCTCAAAATGGTTAGACAAATACAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
C3        TCACCTCTTCTCAAAATGGTTAGACAAATACAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTAAAGGA
CH3       TCACCTCTTCTCAAAATGGTTAGACAAATACAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTAAAGGA
CH11      TCACCTCTTCTCAAAATGGTTAGACAAATACAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTAAAGGA
S11       TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
Z11       TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
S1        TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
Z2        TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
TR        TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
S26       TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
S45       TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
C8        TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
C10       TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
PT12      TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
Z7        TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGCTAACATGACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTCGAGGA
2J        TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
20M       TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
CH2       TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
CH20      TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
CH18      TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
CH8       TCACCTCTTCTCAAAATGGTTAGACAAACGCAACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGAACCGAGACGAATAAATATAACCTTGATAGGTGTTAGAGGA
CH21      TCACCTCTTCTCAAAATGGTTAGACAAACACGACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTGGAGGA
NA        TCACCTCTTCTCAAAATGGTTAGACAAACACGACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTGGAGGA
W4        TCACCTCTTCTCAAAATGGTTAGACAAACACGACGATAATAGGTGGTATAATGTTAACATAACGAAATCACCAGGACCGAGACGAATAAATATAACCTTGATAGGTGTTGGAGGA
con       TCACCTCTTCTCAAAATGGTTAGACAAA--C-ACGAT-ATAGGTGGTATAA-G-TAACAT-ACGAAATCACCAG-ACCGAGACGAATAAATATAAC-TTGATAGGTGTT--AGGA
```

Figure 3.7 Nucleotide sequence alignment of UL146 G1 sequences

All G1 sequences from the present study are included, plus those available in GenBank (in italics). Completely conserved residues are shown in the consensus row (con) and non-conserved residues are indicated by hyphens. Dots indicate gaps in the alignment. Mismatched residues are highlighted in yellow. Sequences encoding predicted signal peptides are highlighted in grey, and those encoding the conserved RCXC motif and two cysteine residues are highlighted in pink.

To illustrate further the high level of sequence conservation within each genotype, an aa sequence alignment of all UL146 G1 strains is shown in Figure 3.6. The sequences are highly conserved, which is in agreement with the high overall identity value for this genotype (97.1%).

Out of a total of 117 aas, only 17 (excluding deletions) differ between the 33 sequences. Three sequences (W4, NA and CH21) also contain two deletions, indicated by dots in their sequences. Figure 3.7 shows a nucleotide sequence alignment of the UL146 G1 strains. All nucleotide sequences within this genotype are highly conserved. Out of 351 nucleotides, only 24 (excluding deletions) are not completely conserved in the 33 sequences.

3.4 UL139 genotypes

A total of 184 DNA samples were tested for the presence of UL139. UL139 was amplified successfully as a PCR product from 168 samples, and sequences were determined successfully from 131 PCR products (Table 3.1). Some samples contained more than one sequence, and in total the 131 samples yielded 183 UL139 sequences.The phylogenetic and diversity analyses involved a total of 300 UL139 sequences, which included the 183 sequences derived from the present study (Table 3.1) and all other UL139 sequences reported by others in the literature or deposited in GenBank (Table 3.4).

Table 3.4: Additional UL139 Sequences Used for Phylogenetic and Diversity Analyses

Accession numbers	*Reference*
AY905263, AY905264, AY601874-AY601877, AY218873-AY218887	Qi *et al.* (2006)
DQ180386, DQ180358, DQ180374	Zhou *et al.* (unpublished)
AY999242-AY999271, AY805250-AY805303, AY818255-250	Mao *et al.* (unpublished)
DQ229942-DQ229948	Ruan *et al.* (unpublished)

The UL139 coding sequences range in length from 372-444 bp (124-148 codons, including the stop codon). The aa sequence alignments and phylogenetic trees produced by the neighbour-joining method were employed to group the sequences into genotypes. All of the sequences fell into eight UL139 genotypes (G1-G8). Table 3.5 displays a laboratory strain or previously published strain (Dolan *et al.,* 2004) representing each of the eight genotypes. An aa sequence alignment containing a representative of each genotype is shown in Figure 3.8A. The protein encoded by each HCMV UL139 genotype contains a putative signal peptide sequence and a transmembrane region (both highlighted in grey).

Unlike UL146, variation in UL139 is concentrated in the N-terminal region of the protein. Indeed, the C terminus is highly conserved between genotypes. An unrooted phylogenetic tree showing the relationship between the eight UL139 genotypes and CCMV UL139 is shown in Figure 3.8B. Bootstrap values under 70 indicate regions of unresolved branching order. G2 and G7 appear to be a related pair of genotypes, whereas G1 and G5 appear distant from the other UL139 genotypes.

Table 3.5: Laboratory strain corresponding to UL139 Genotype

UL146 Genotype	*Laboratory strain*
G1	Merlin
G2	JP
G3	NT
G4	Toledo
G5	Towne
G6	3157
G7	W
G8	A3*

* No Laboratory strain fell into G8

Genotypic frequencies were calculated from the total number of UL139 sequences available (from the present study and in Genbank, a total of 300)

(Figure 3.9). UL139 G2 and G4 were each detected in over 22% of sequences, and UL139 G1 in more than 15%. In contrast, some genotypes were detected at very low frequencies, for example G7 and G8, being identified in less than 5% sequences. The other three genotypes (G3, G5 and G6) were detected at frequencies of 7-10%.

The aa sequence alignments were used to calculate the level of sequence identity between all strains within each individual genotype (using Swaap) (Figure 3.10). An aa sequence alignment of a representative of each UL139 genotype (Figure 3.8A) was used to calculate identity among genotypes. The aa sequence identity among genotypes is low (range 54.9-97.2%, mean 80.15%), whereas within each genotype it is high (range 85.9-98.6%). Variation within genotypes tends to be higher in UL139 than in UL146, but lower among genotypes. Similarly, nucleotide sequence identity is low between genotypes (range 64.6-97.2%, mean 84.3%), whereas it is high within genotypes (range 94.6-99.2%). Sequences in UL139 G1 exhibit a greater level of variation than those in the other UL139 genotypes (94.5% at the nucleotide sequence level, 90% at the aa sequence level). This is due to a small number of strains in G1 that may represent a subgenotype. Therefore despite the unresolved branching order, the clustering of sequences into eight groups with high sequence identity confirm the presence of eight genotypes.

To illustrate further the high level of sequence conservation within each genotype, an aa sequence alignment of the 29 UL139 G3 sequences is shown in Figure 3.11. As some of the sequences obtained did not include the C terminus, this region was removed from all sequences in the set. The sequences are highly conserved, which is in agreement with the high overall identity value for this genotype (98.6%).

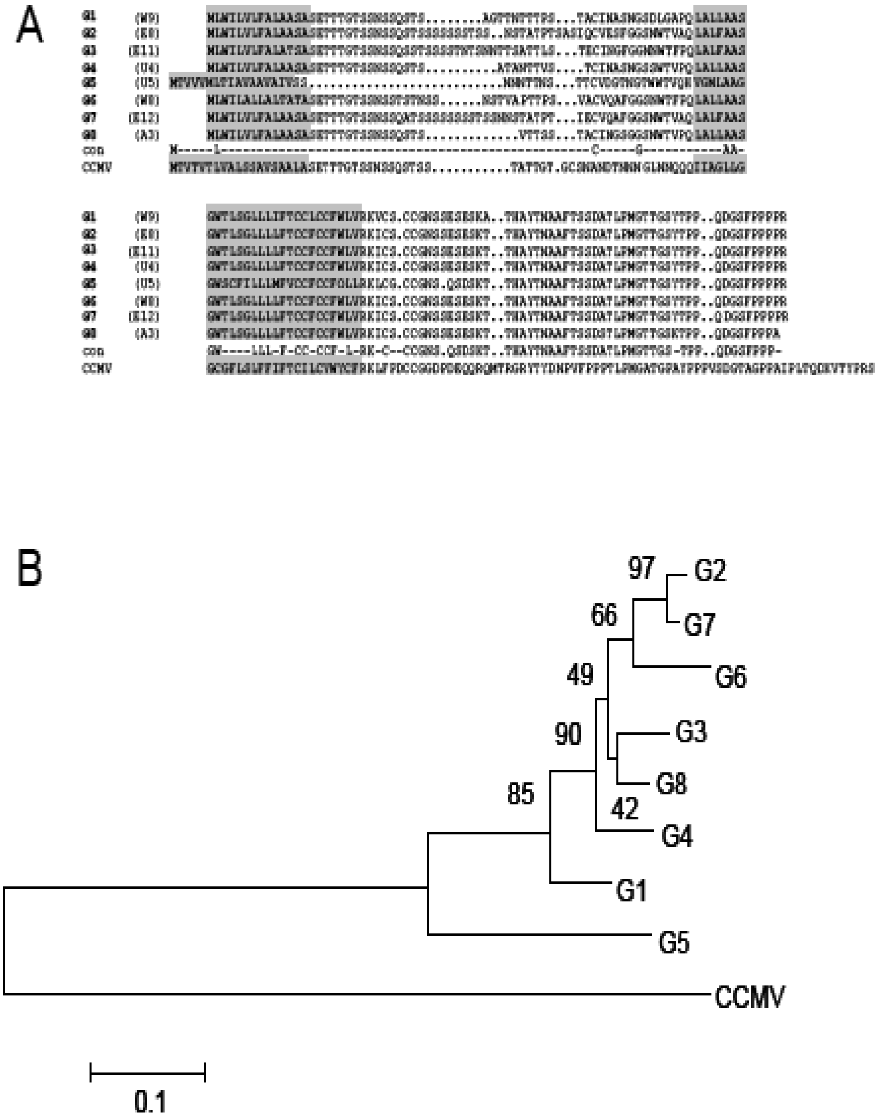


Figure 3.8 Amino acid sequence alignment and phylogenetic analysis of UL139 genotypes

A) An alignment of amino acid sequences representing the eight genotypes (G1-G8), with representative strains (Table 3.1) indicated in parentheses. Completely conserved residues are shown in the consensus row (con) and non-conserved residues are represented by hyphens. The (partial) CCMV UL139 sequence is shown below the consensus. Predicted signal sequences and transmembrane regions are highlighted in grey.

B) The unrooted neighbour-joining phylogenetic tree of representatives of the eight UL139 genotypes (G1-G8) and CCMV UL139 was produced in Mega 4.0, using the amino acid sequence alignment shown in Figure 3.8A. The scale bar indicates divergence as substitutions per amino acid site. Bootstrap values are out of 100.

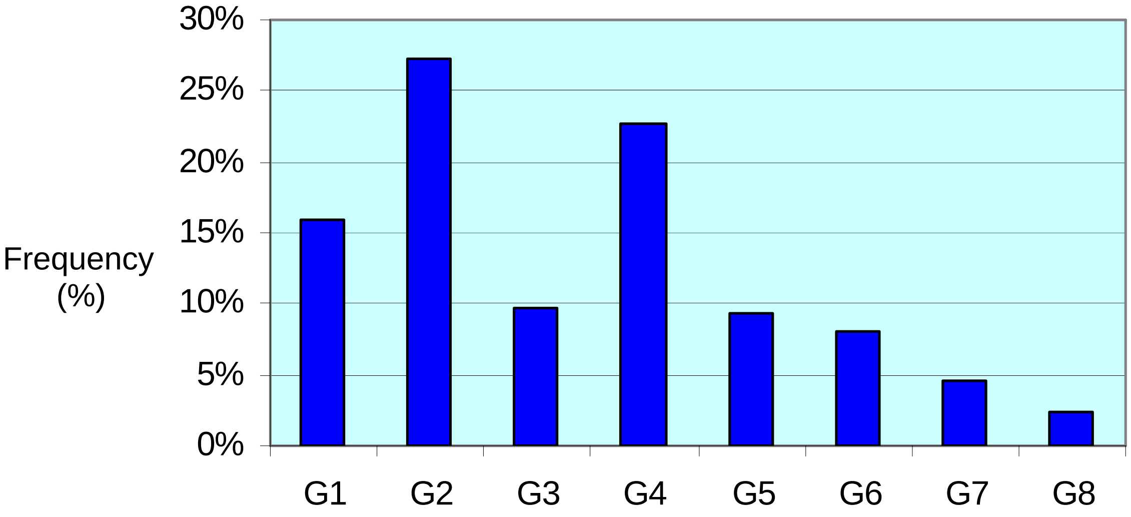


Figure 3.9 Frequencies of occurrence of UL139 genotypes
The frequencies were calculated using all available sequences from the present study and those available in GenBank (a total of 300).

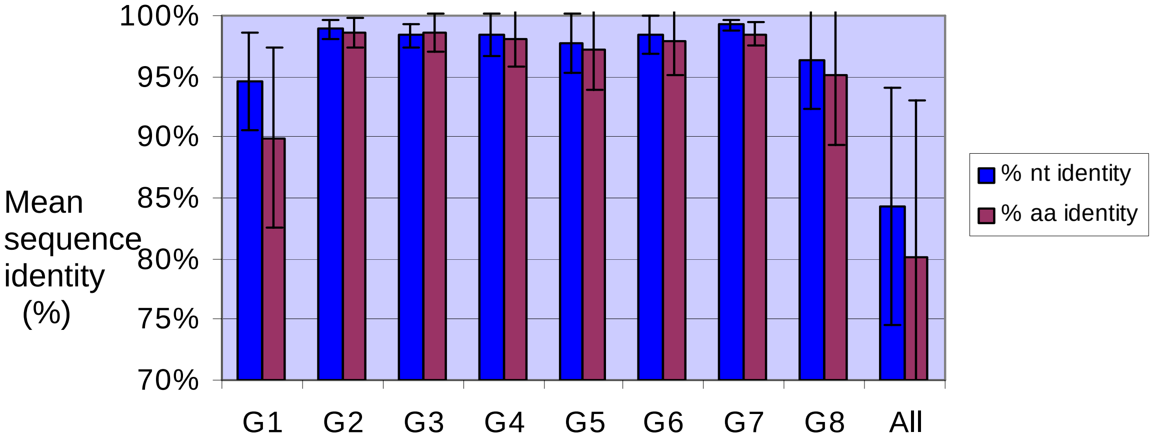


Figure 3.10 Sequence identity within and among UL139 genotypes

Nucleotide (nt) and amino acid (aa) sequence identities were calculated *within* each genotype (G1-G8) by pairwise alignment of all sequences (these were partial sequences as they lacked 29 amino acid-encoding codons from the highly conserved C terminus) in the relevant genotype, and *among* all genotypes (All) by pairwise alignment of a representative of each genotype (Figure 3.8A). Mean and standard deviation values are shown.

Out of a total of 110 aas, only 10 (excluding deletions) differ between the sequences. Two sequences (D5 and N2) contain single deletions. Figure 3.12 shows a nucleotide sequence alignment of the UL139 G3 sequences. Out of 330 nucleotides, only 22 (excluding deletions) differ between the 29 sequences.

3.5 CCMV UL139

CCMV UL139 is much larger (359 codons) than HCMV UL139 (124-148 codons) (Davison *et al.,* 2003). The predicted aa sequence of CCMV UL139 is related to two proteins: to HCMV UL139 in its C terminal region (shown in Figure 3.8A) and to the rhesus cytomegalovirus (RhCMV) protein encoded by rh174 gene in its N terminal region. Figure 3.13 shows an aa sequence alignment of the N-terminal portion of CCMV UL139 with the whole of rh174 and the C-terminal portion of CCMV UL139 with the whole of HCMV UL139 (G1). CCMV UL139 contains three hydrophobic regions: a signal sequence at the N-terminus that corresponds to that of rh174, an internal region that corresponds to the signal sequence of HCMV UL139, and a second internal region that corresponds to the transmembrane anchor of HCMV UL139.

3.6 Mode of selection of UL146 and UL139

A nucleotide substitution that results in an aa change is called a non-synonymous substitution, and a substitution that does not result in an aa change is termed a synonymous substitution. To assess the mode by which UL146 and UL139 have evolved, the frequencies of non-synonymous substitution (dN) and synonymous substitution (dS) were investigated. If a gene has evolved under neutral selection with non-synonymous and synonymous sites having evolved at equal rates, then dN=dS (i.e. dN/dS=1). This is often called the null hypothesis. If a gene has evolved under positive selection with non-synonymous sites having evolved faster than synonymous sites, then dN>dS (i.e. dN/dS>1).

```
D5              MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
N2              MLWILVLFALATSASETTTGTSSNSSQSST.SSSTNTSNNTTSATTLSTECINGFGGN
CH10M           MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH13J           MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH14J           MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH15J           MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH175           MLWILILFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH181           MLWILVLFALATSASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH18M           MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH194           MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSVTTLSTECINGFGGN
CH27C           MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSVTTLSTECINGFGGN
CH282           MLWILVLFALATSASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH290           MLWILILFVLAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH291           MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH29C           MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSVTTLSTECINGFGGN
CH38M           MLWILVLFALATSASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH41M           MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH45J           MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH63J           MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH83            MLWILILFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
CH88            MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
S37             MLWILVLFALAASASETTTGTSSKSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
E11             MLWILVLFALATSASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
G2              MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
C2              MLWILILFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
C3              MLWILILFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
C7              MLWILVLFALAASASETTTGTSSNSSQSSTSSSSTNTSNNTISATTLSTECINGFGGN
C8              MLWILVLFALAASASETNTGTSYNSSQSSTSSSSTNTSNNTISATTLSTECINGFGGN
E12             MLWILVLFALATSASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLSTECINGFGGN
con             MLWIL-LF-LA-SASET-TGTS--SSQSST-SSSTNTSNNT-S-TTLSTECINGFGGN

D5              NWTFPQLALFAASGWTLSGLLLLLTCCFCCFWLVRKICSCCGN.SESESKTT
N2              NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH10M           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH13J           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH14J           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH15J           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH175           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH181           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH18M           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH194           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH27C           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH282           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH290           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH291           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTN
CH29C           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH38M           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH41M           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH45J           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH63J           NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH83            NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
CH88            NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
S37             NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
E11             NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
G2              NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
C2              NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
C3              NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
C7              NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
C8              NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWIVRKICSCCGNSSESESKTT
E12             NWTFPQLALFAASGWTLSGLLLLFTCCFCCFWLVRKICSCCGNSSESESKTT
con             NWTFPQLALFAASGWTLSGLLLL-TCCFCCFWLVRKICSCCGN-SESESKT-
```

Figure 3.11 Amino acid sequence alignment of partial UL139 G3 sequences (continued overleaf)

Figure 3.11 Amino acid sequence alignment of partial UL139 G3 sequences

All G3 sequences from the present study are included, plus those available in GenBank (in italics). As only the first 110 aa were obtained for some sequences, C terminal residues have been trimmed. Completely conserved residues are shown in the consensus row (con) and non-conserved residues are indicated by hyphens. Dots indicate gaps in the alignment. Mismatched residues are highlighted in yellow. The predicted signal peptide sequences and transmembrane regions are highlighted in grey.

```
E11     ATGCTGTGGATATTAGTTTTATTTGCACTCGCCACATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
N2      ATGCTGTGGATATTAGTTTTATTTGCACTCGCCACATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTC...CACTAATACCAGTAA
E12     ATGCTGTGGATATTAGTTTTATTTGCACTCGCCACATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH88    ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH13J   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH18M   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH14J   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH41M   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH15J   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
G2      ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH45J   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
C2      ATGCTGTGGATATTAATTTTATTTGCACTCGCCGCATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH290   ATGCTGTGGATATTAATTTTATTTGTACTCGCCGCATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
C3      ATGCTGTGGATATTAATTTTATTTGCACTCGCCGCATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH175   ATGCTGTGGATATTAATTTTATTTGCACTCGCCGCATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH282   ATGCTGTGGATATTAGTTTTATTTGCACTCGCCACATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
D5      ATGCTGTGGATATTAGTTTTATTTGCACTCGCCGCATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH83    ATGCTGTGGATATTAATTTTATTTGCACTCGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH291   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH29C   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH194   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH27C   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH63J   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH10M   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH181   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCACATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
CH38M   ATGCTGTGGATATTAGTTTTATTTGCACTTGCCACATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
S37     ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAAGTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
C7      ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
C8      ATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCAAGTGAAACCAATACAGGTACCAGCTATAATTCCAGTCAATCATCTACTTCATCTTCTTCCACTAATACCAGTAA
con     ATGCTGTGGATATTA-TTTTATTTG-ACT-GCC-CATCGGC-AGTGAAACCA-TACAGGTACCAGCT-TAA-TCCAGTCAATCATCTACTTCATCTTC---CACTAATACCAGTAA

E11     TAATACCACAAGTGCAACCACACTGTCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
N2      TAATACCACAAGTGCAACCACACTGTCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
E12     TAATACCACAAGTGCAACCACACTGTCAACAGAATGCATTAATGGTTTTGGTGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH88    TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH13J   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH18M   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH14J   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH41M   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH15J   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
G2      TAATACCACAAGTGCAACCACACTGTCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH45J   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
C2      TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH290   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
C3      TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH175   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH282   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
D5      TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH83    TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH291   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH29C   TAATACCACAAGTGTAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH194   TAATACCACAAGTGTAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH27C   TAATACCACAAGTGTAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH63J   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH10M   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH181   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
CH38M   TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
S37     TAATACCACAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
C7      TAATACCATAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
C8      TAATACCATAAGTGCAACCACACTATCAACAGAATGCATTAATGGTTTTGGCGGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
con     TAATACCA-AAGTG-AACCACACT-TCAACAGAATGCATTAATGGTTTTGG-GGCAATAATTGGACATTTCCACAACTCGCGCTGTTTGCTGCTAGCGGCTGGACATTATCTGGAC
```

Figure 3.12 Nucleotide sequence alignment of partial UL139 G3 sequences (continued overleaf)

```
E11         TCCTTCTCTTATTTACCTGCTGCTTTTGCTGCTTTTGGCTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCAGAGTCAGAGAGCAAAACAACC
N2          TCCTTCTCTTATTTACCTGCTGCTTTTGCTGCTTTTGGCTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCAGAGTCAGAGAGCAAAACAACC
E12         TCCTTCTCTTATTTACCTGCTGCTTTTGCTGCTTTTGGCTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCAGAGTCAGAGAGCAAAACAACC
CH88        TCCTTCTCTTATTTACCTGCTGCTTTTGCTGCTTTTGGCTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
CH13J       TCCTTCTCTTATTTACCTGCTGCTTTTGCTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACT
CH18M       TCCTTCTCTTATTTACCTGCTGCTTTTGCTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACT
CH14J       TCCTTCTCTTATTTACCTGCTGCTTTTGCTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACT
CH41M       TCCTTCTCTTATTTACCTGCTGCTTTTGCTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACT
CH15J       TCCTTCTCTTATTTACCTGCTGCTTTTGCTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACT
G2          TCCTTCTCTTATTTACCTGCTGCTTTTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
CH45J       TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
C2          TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCGGAGAGCAAAACAACC
CH290       TCCTTCTTTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCGGAGAGCAAAACAACC
C3          TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCGGAGAGCAAAACAACA
CH175       TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCGGAGAGCAAAACAACC
CH282       TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCGGAGAGCAAAACAACC
D5          TCCTTCTCTTACTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACT...CCGAGTCAGAGAGCAAAACAACC
CH83        TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
CH291       TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCGGAGAGCAAAACAAAC
CH29C       TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
CH194       TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
CH27C       TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
CH63J       TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
CH10M       TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
CH181       TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
CH38M       TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
S37         TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
C7          TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGTTAGTACGTAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
C8          TCCTTCTCTTATTTACCTGCTGCTTCTGTTGCTTTTGGATAGTACGCAAAATCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCAAAACAACC
con         TCCTTCT-TTA-TTACCTGCTGCTT-TG-TGCTTTTGG-TAGTACG-AAAATCTGCAGCTGCTGCGGCAACT---C-GAGTC-GAGAGCAAAACAA--
```

Figure 3.12 Nucleotide sequence alignment of partial UL139 G3 sequences

All G3 sequences from the present study are included, plus those available in GenBank (in italics). The alignment shows the first 330 nucleotides. Completely conserved residues are shown in the consensus row (con) and non-conserved residues are indicated by hyphens. Dots indicate gaps in the alignment. Mismatched residues are highlighted in yellow. Sequences encoding predicted signal peptides and transmembrane regions are highlighted in grey.

```
CCMV UL139      MTEL.....RLASSLLLAW......EILDMSLANSYTQGP....KIHANV
RhCMV rh174     MMQPVSHKHAHALILLMMWPSGTSGEALRCA.SKDYTSNTDMKYQLPINM
con             M----------A—-LL--W------E-L-------YT-----------N-

CCMV UL139      TKCI.WKVKNETLVIGKYIDIKCEFTEVTDNITMGLMYTSCSNRSLSETT
RhCMV rh174     TKVVSGPTKNETLS.GPYT.VTCIF..VSGECTRGIYFTACDQNATTQVY
con             TK------KNETL—-G-Y----C-F—-V----T-G---T-C---------

CCMV UL139      MSYYNHTPQDRANYPFGKVEHSRSDTTATMTLRRCGLNCTGIHDCFKFDG
RhCMV rh174     M.YKNHAP....IFPNDESKTSASNTNGFTMRWPVSPHAPGTYDCFSYDN
con             M-Y-NH-P------P------S-S-T--------------G--DCF---D

CCMV UL139      --QKMNITLRTSIAPIGTIYVNTKA.NQS.HDVFCAVNDTFPATVLLHNT
RhCMV rh174     VTNIMNITQRTQVTPMGITYASKHSTNQSVYNVSCSFNSTFPGTVTLIVQ
con             ----MNIT-RT---P-G--Y------NQS---V-C—-N-TFP-TV-L---

CCMV UL139      GNDHQLNITTNHTVKKCNRTLYYGYTV..QGISPPTQCSLFSSTCI....
RhCMV rh174     GAVNA.TVIHNETVKLCGQDLYWNYLVLTSGGTPTFQCTNKATNCLLSGY
con             G---------N-TVK-C---LY—-Y-V---G--P—-QC------C-----

CCMV UL139      .GLRSHTLTYPGTPLTPPAGISNCENYTAPN*
RhCMV rh174     SRLWSNTTSTPGPPIPI...LHNCSTYSHPWWTTARPVTTPSVSTTQLTS
con             --L-S-T---PG-P--------NC--Y--P

RhCMV rh174     LSISASTSFTYNVSLVVYEAQYASRELHGLWILVVLIICAAVACWLRLPQ

RhCMV rh174     VVVQMFRKCVASLQRKHNVYTNM


CCMV UL139      MTVTVTLVALSSAVSAALASETTTGTSSNSSQSTS....STATTGTGCSN
HCMV UL139G1    MLWILVLFALAAS.....ASETTTGTSSNSSQSTSAGTTNTTTPSTACIN
con             M-------AL--------ASETTTGTSSNSSQSTS-----T-T--T-C-N

CCMV UL139      ANDTNNNGLNQQQIIAGLLGGCGFLSLFFIFTCILCVWYCFRKLFPDCCG
HCMV UL139G1    ASNGSDLGAPQLALLAA..SGWTLSGLLLIFTCCLCCFWLVRKVCS.CCG
con             A-N----G--Q----A----G-----L--IPTC-LC-----RK----CCG

CCMV UL139      GDPDEQQRQMTRGRYTYDNPVFPPP..TLPMGATGPAYPPPVSDGTAGPP
HCMV UL139G1    NSSESE....SKATHAYTNAAFTSSDATLPMGTTG.SYTPPQDGSFPPPP
con             ----------------Y-N--F-----TLPMG-TG--Y-PP-------PP

CCMV UL139      AIPLTQDKVTYPRS
HCMV UL139G1    ............R.
con             ------------R-
```

Figure 3.13 Amino acid sequence alignment of CCMV UL139 with RhCMV rh174 and HCMV UL139

A) The N-terminal portion of CCMV UL139 aligned with the whole of RhCMV rh174 (accession number NC_006150). The residue marked by an asterisk in (A) immediately precedes the first CCMV UL139 residue in (B). B) The C-terminal portion of UL139 CCMV aligned with the whole of HCMV UL139 (G1). Predicted signal sequences are highlighted in grey and predicted transmembrane regions are highlighted in yellow. Gaps in the alignment are represented by dots. Completely conserved residues are shown in the consensus row (con) and non-conserved residues are represented by hyphens.

If a gene has evolved under selective constraint (or purifying selection) with non-synonymous sites having evolved more slowly than synonymous sites, then dN<dS (i.e. dN/dS<1) (Nei and Gojobori, 1986; reviewed by Hurst, 2002).

Intragenotypic dN/dS values were calculated from nucleotide alignments of all sequences in each UL146 (Table 3.6) and UL139 genotype (Table 3.7), using Swaap 1.0.2 (method of Nei and Gojobori, 1986). For UL146, dN/dS>1 was obtained for G1, G2, G4, G13 and G14. An intergenotypic dN/dS value was calculated for a nucleotide alignment of a representative strain of each UL146 genotype and was <1 (Table 3.6). UL139 intragenotypic dN/dS values were calculated using a trimmed nucleotide alignment (first 330 nucleotides), as the reverse primer used to amplify some sequences is located with the UL139 coding region. For all UL139 genotypes, intragenotypic dN/dS values were <1 (Table 3.7). Similarly, the intergenotypic dN/dS value for UL139 was <1. Representative strains for each UL146 and UL139 genotype were those used previously (Figures 3.2 and 3.8).

Table 3.6: UL146 Diversity

Genotype	*Samples*	*Frequency*	*Identity*[a]		*dN/dS*[b]	*Z-test*[d]
		%	DNA %	Protein %		
G1	34	9.71	98.82	97.11	2.21	P (*p*=0.068)
G2	25	7.14	99.76	99.48	1.15	N (*p*=0.473)
G3	10	2.86	99.05	98.43	0.38	N (*p*=1)
G4	8	2.29	99.57	98.92	1.12	N (*p*=0.384)
G5	16	4.57	99.31	98.79	0.37	N (*p*=1)
G6	2	0.57	97.15	94.87	0.38	N (*p*=1)
G7	57	16.3	98.96	97.91	0.54	N (*p*=1)
G8	22	6.29	99.41	99.16	0.22	N (*p*=1)
G9	49	14	98.36	96.79	0.49	N (*p*=1)
G10	12	3.43	99.59	99.53	0.19	N (*p*=1)
G11	19	5.43	99.43	98.84	0.57	N (*p*=1)
G12	43	12.3	98.44	98.15	0.16	C (*p*=1)
G13	47	13.4	99.29	98.37	4.94	P (*p*=0.042)
G14	6	1.71	99.29	98.37	4.93	P (*p*=0.029)
All	350	100	59.29	35.96	0.82[c]	P (*p*=1)

[a] Mean nucleotide and aa identity from Swaap 1.0.2.
[b] Average dN/dS from Swaap 1.0.2.
[c] Calculated from a comparison of a single member of each genotype.
[d] Codon-based Z-test from Mega 4.0, N is neutrality, P is positive selection, and C is constraint. *p* is the probability that this genotype is evolving under positive selection

In addition to this overall analysis, dN and dS values were calculated by pairwise comparison of all strains within each UL146 and UL139 genotype, using Swaap and the method of Nei and Gojobori (1986). Values of dN and dS for each pair of sequences were plotted on a scatter plot for each genotype, as shown in Figures

3.14 and 3.15. A diagonal line where dN/dS=1 was also plotted. When the majority of points are above the diagonal this indicates an excess of synonymous substitutions, and when the majority of points are below the diagonal this indicates an excess of non-synonymous substitutions.

Table 3.7: UL139 Diversity

Genotype	*Samples*	*Frequency*	*Identity*[a]		*dN/dS*[b]	*Z-test* [d]
		%	DNA %	Protein %		
G1	48	16	94.56	89.97	0.57	N (*p*=1)
G2	82	27.33	98.87	98.56	0.18	C (*p*=1)
G3	29	9.66	98.37	98.57	0.12	C (*p*=1)
G4	68	22.66	98.43	98.15	0.22	C (*p*=1)
G5	28	9.33	97.71	97.19	0.19	C (*p*=1)
G6	24	8	98.49	97.96	0.30	N (*p*=1)
G7	14	4.66	99.22	98.51	0.66	N (*p*=1)
G8	7	2.33	96.37	95.19	0.46	N (*p*=1)
All	300	100	86.17	82.51	0.35[c]	C (*p*=1)

[a] Nucleotide and aa identity from Swaap 1.0.2.
[b] dN/dS from Swaap 1.0.2.
[c] Calculated from a comparison of a single member of each genotype.
[d] Codon-based Z-test from Mega 4.0, N is neutrality and C is constraint.
p is the probability that this genotype is evolving under positive selection

For UL146 G1 there are equal numbers of points above and below the diagonal (and one point on the line), indicating dN=dS (Figure 3.14). For G5, G6, G7, G8, G11, G12, G13 and G14 the majority of points are above the diagonal, and for G2, G3, G4, G9 and G10 the majority of the points are below the diagonal. For all UL139 genotypes the majority of points are above the diagonal (Figure 3.15). Values of dN and dS were also calculated by pairwise comparison of a representative strain of each UL146 genotype and plotted on a scatter plot. The majority of points are above the diagonal (Figure 3.16). Similarly, values of dN and dS were calculated by pairwise comparison of a representative strain of each UL139 genotype and all points are above the diagonal (Figure 3.16).

Nucleotide sequence alignments of all strains in each UL146 and UL139 genotype were analysed further by performing a codon based Z-test using Mega 4.0, where the probability *p* of rejecting the null hypothesis (dN/dS=1) in favour of an alternative theory (i.e. positive selection, dN/dS>1 or constraint, dN/dS<1) was calculated (Nei and Kumar, 2000). The *p* values shown are the probability of rejecting the null hypothesis in favour of positive selection. This is a statistical test that compares the frequencies of dS and dN between sequences. Values of *p*<0.05 are significant at the 5% level, whereas *p*<0.01 are significant at the 1% level.

Results from the Z-test are shown in Tables 3.6 and 3.7, and indicate whether a particular genotype scored is likely to have evolved under neutrality, constraint or positive selection. From this analysis, ten UL146 genotypes (G2-G11) scored as having evolved under neutrality, one (G12) under constraint, and three (G1, G13 and G14) under positive selection, although the results were significant at the 5% level only for G13 and G14. Four of the eight UL139 genotypes (G2-G5) scored as having evolved under constraint and four (G1, G6, G7 and G8) under neutrality. A codon based Z-test was also performed on a nucleotide sequence alignment of a representative of each UL146 genotype, and UL146 genotypes scored as having evolved under positive selection, although this was not statistically significant. From the Z-test, UL139 genotypes scored as having evolved under constraint.

The *codeml* program from the PAML package was used for further analysis of all sequences in the genotypes that showed some evidence for positive selection: UL146 G1, G2, G3, G4, G9, G10 and G13. This program carries out maximum likelihood analysis of protein-coding DNA sequences using codon substitution models (Goldman and Yang 1994; Suzuki and Gojobori, 1999; Yang *et al.,* 2000; Yang and Nielsen 2002). Values of dN/dS>1 were obtained for UL146 G1, G2 and G13, but were only significant at the 5% level for G1 (*p*=0.046) and no residues under positive selection were identified.

3.7 Geographical distribution of UL146 and UL139 genotypes

To analyse the geographical distribution of UL146 and UL139 genotypes, sample origin was divided into four continental regions: Africa, Asia, Europe and Australia. American samples were excluded owing to insufficient numbers. Table 3.8 shows the number of sequences in each UL146 genotype for each of the four regions.

There are some examples of apparent geographical isolation of genotypes. For example UL146 G10 and G11 were detected only in European samples, and the single example of UL146 G6 was found in an Asian sample.

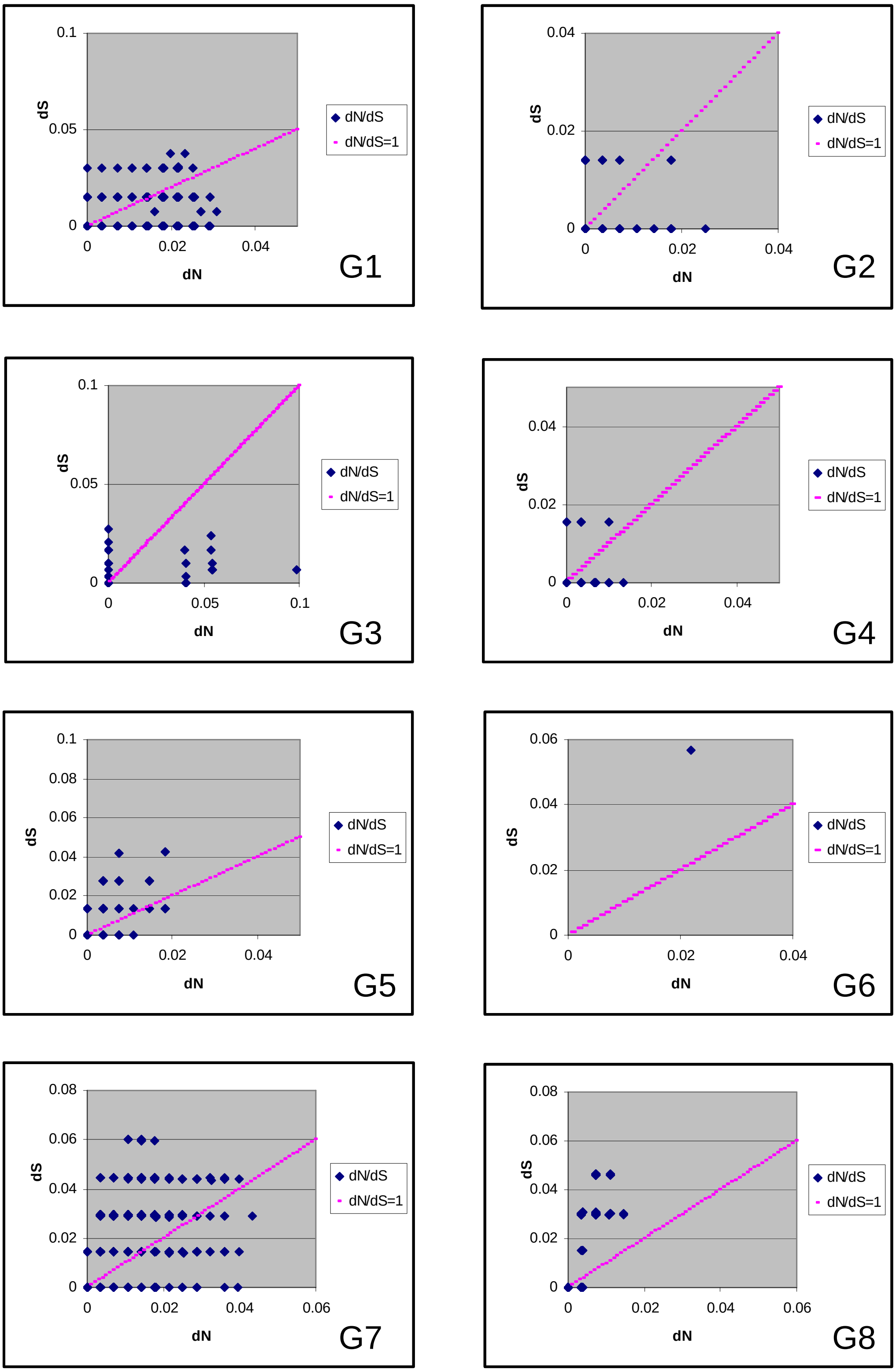


Figure 3.14 Pairwise comparison of dN/dS between all strains in each UL146 genotype (G1-G14) (continued overleaf)

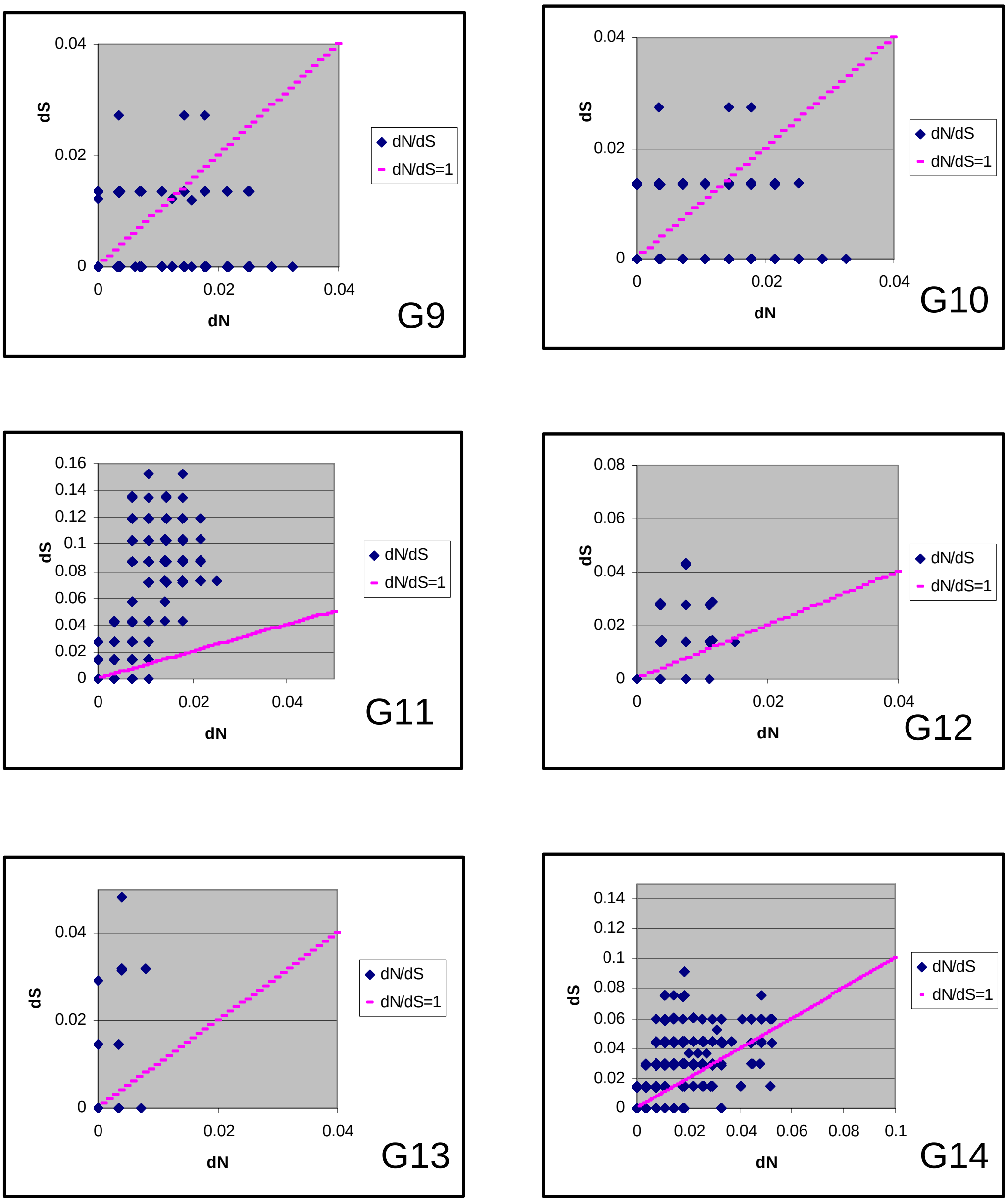


Figure 3.14 Pairwise comparison of dN/dS between all strains in each UL146 genotype (G1-G14)

(G1-G8)

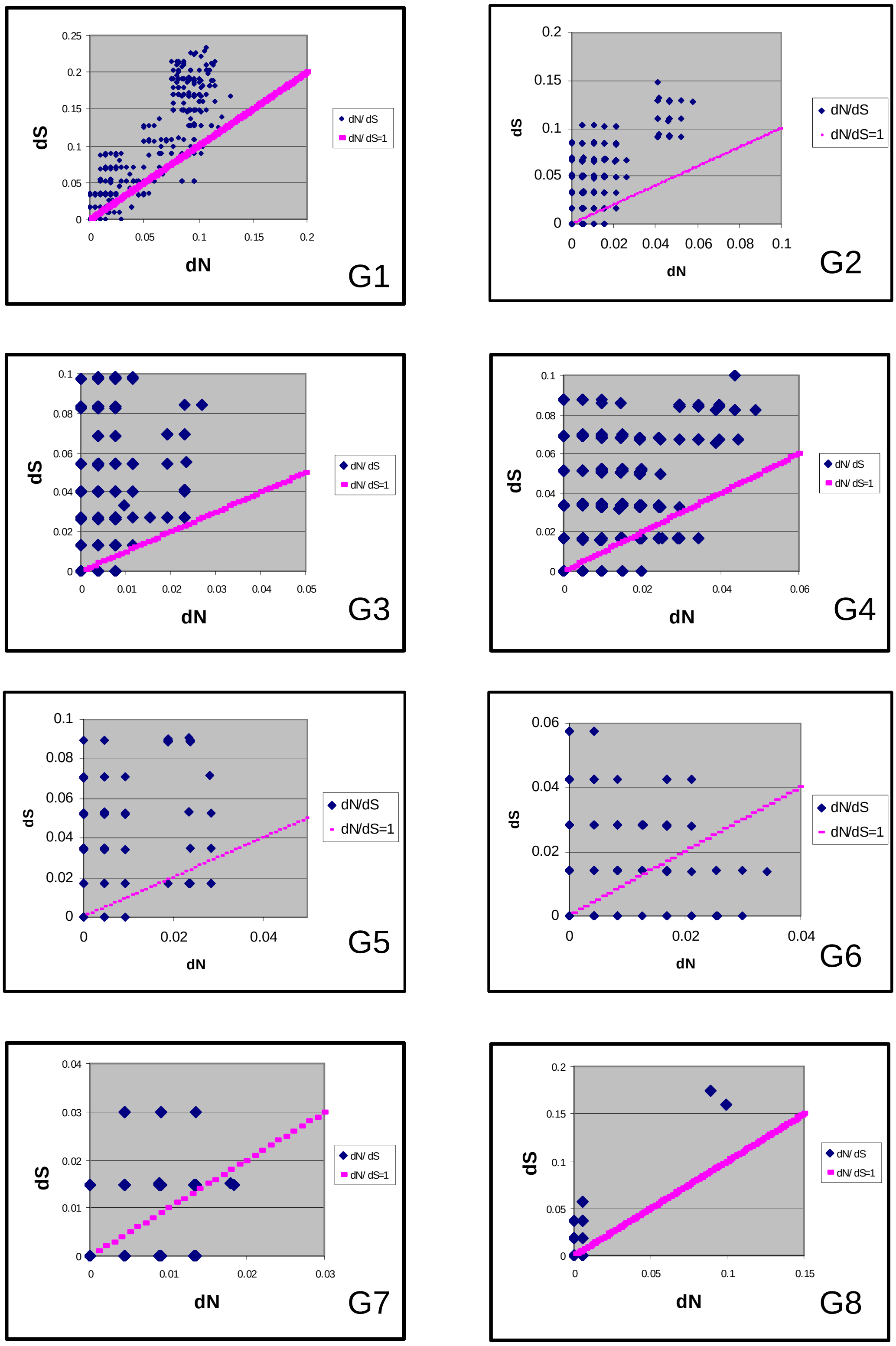
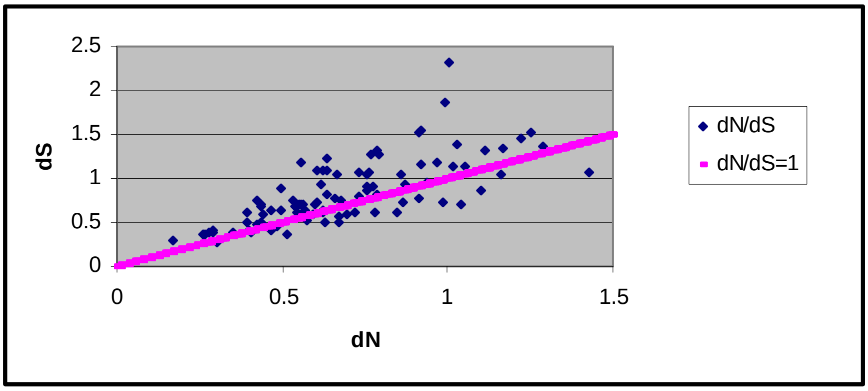
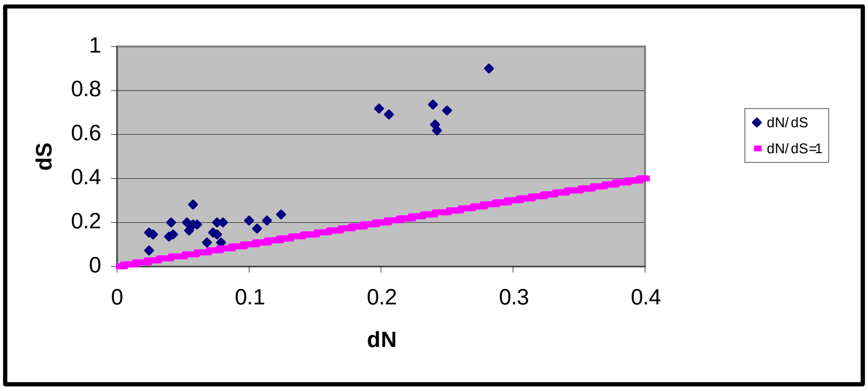


Figure 3.15 Pairwise comparison of dN/dS between all strains in each UL139 genotype (G1-G8)



UL146



UL139

Figure 3.16 Pairwise comparison of dN/dS between a representative strain of each UL146 or UL139 genotype

The data in Table 3.8 are presented visually in Figure 3.17. The European and African samples show similar patterns of genotypic distribution, although G10 and G11 were not found in African samples.

Table 3.8: Geographical Distribution of UL146 Genotypes

Genotype	*Africa*	*Asia*	*Europe*	*Australia*	
G1	4	2	7	0	
G2	2	1	7	2	
G3	4	0	2	0	
G4	1	0	4	2	
G5	3	0	2	1	
G6	0	1	0	0	
G7	5	6	21	1	
G8	2	0	5	0	
G9	6	2	11	3	
G10	0	0	8	0	
G11	0	0	3	0	
G12	5	1	16	0	
G13	11	0	17	6	
G14	2	0	1	0	
Totals	45	13	104	15	177

The data in Table 3.8 were analysed further by a chi-square test in order to assess whether the observed frequencies differ significantly from the expected frequencies (displayed in Table 3.9). The expected frequencies are based on the null hypothesis, which assumes independent, random distribution of all genotypes in all regions. Yate's correction was applied during chi-square analysis to correct for cells with frequencies below 5, however for cells where the frequency is zero, results obtained need to be regarded with caution.

Table 3.9: Statistical Analysis of the Geographical Distribution of UL146 Genotypes

UL146	*p*
G1	0.485
G2	0.722
G3	0.131
G4	0.241
G5	0.391
G6	0.006
G7	0.047
G8	0.723
G9	0.777
G10	0.132
G11	0.551
G12	0.408
G13	0.073
G14	0.422

Chi-square test applied across each row in Table 3.86.

UL146 G6 (*p*=0.006) and G7 (*p*=0.047) show statistically significant differences in their genotypic distributions, whereas UL146 G10 and G11 do not.

Table 3.10 shows the number of sequences in each UL139 genotype for each of the four regions. No geographical isolation of UL139 genotypes was detected. The data in Table 3.10 are presented visually in Figure 3.18. The genotypic distribution patterns for Africa, Asia and Europe are similar, although G6 was not found in African samples. The data in Table 3.10 were analysed further by a chi-square test, and the results are shown in Table 3.11. UL139 G3 (*p*=0.032) and G7 (*p*=0.006) show statistically significant differences in their genotypic distributions.

Table 3.10: Geographical Distribution of UL139 Genotypes

Genotype	*Africa*	*Asia*	*Europe*	*Australia*	
G1	8	2	16	2	
G2	9	1	23	5	
G3	1	4	7	0	
G4	8	5	23	5	
G5	10	3	14	0	
G6	0	1	11	1	
G7	3	0	5	5	
G8	2	2	1	1	
Totals	41	18	100	19	178

Table 3.11: Statistical Analysis of the Geographical Distribution of UL139 Genotypes

UL139	*p*
G1	0.814
G2	0.483
G3	0.032
G4	0.921
G5	0.151
G6	0.168
G7	0.006
G8	0.148

Chi-square test applied across each row in Table 3.10.

Identical nucleotide sequences were frequently obtained from geographically and, presumably, epidemiologically unrelated patients.

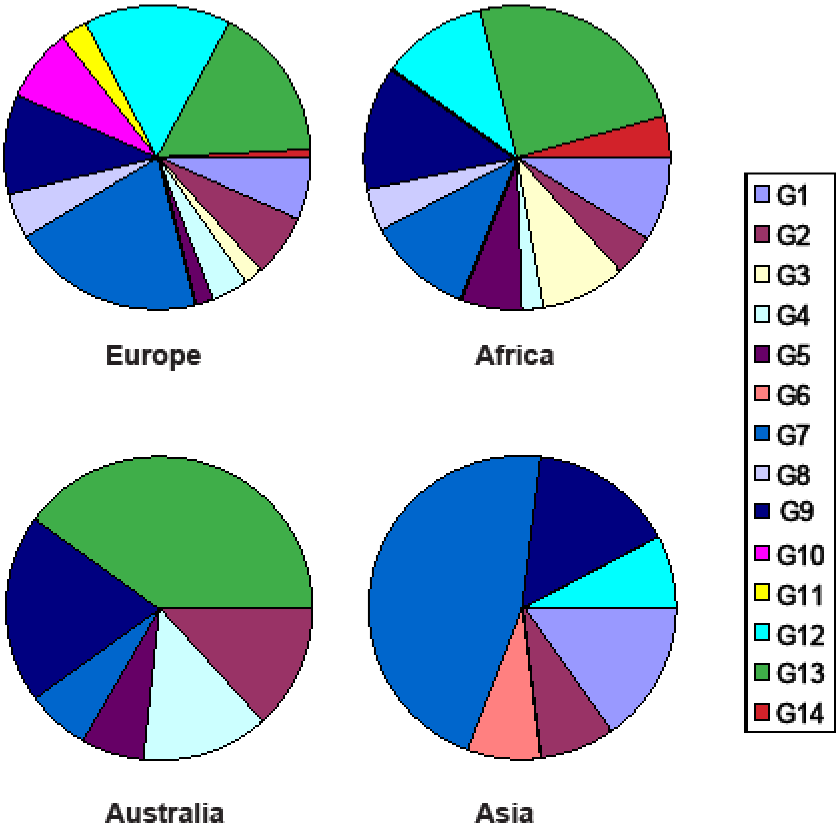


Figure 3.17 Geographical distribution of UL146 genotypes

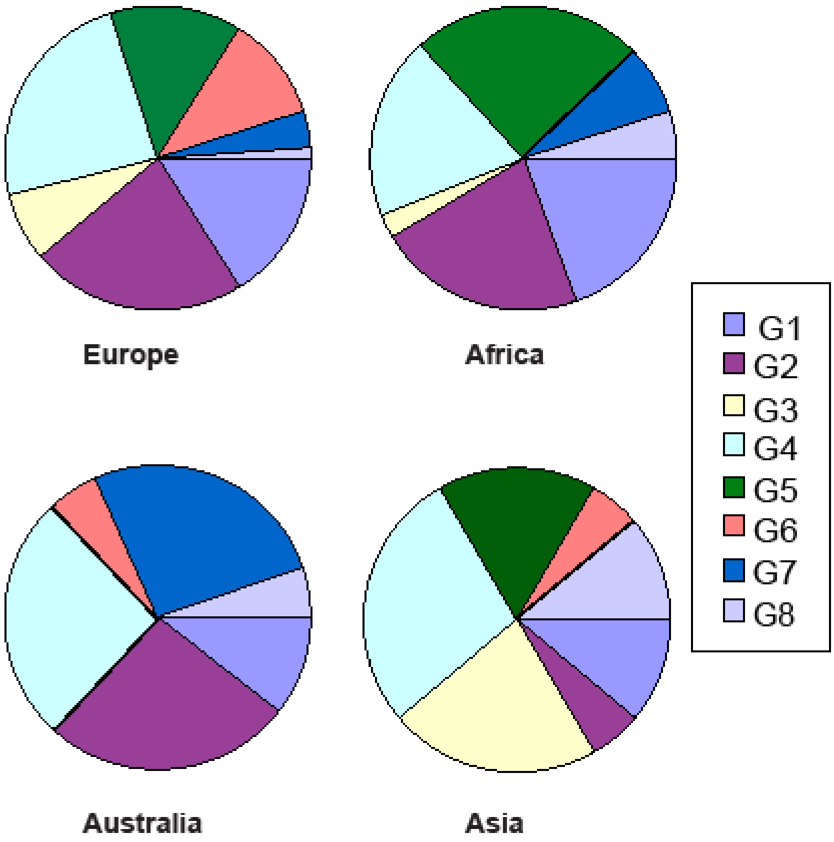


Figure 3.18 Geographical distribution of UL139 genotypes

For example, certain samples from The Gambia, Scotland and Hungary contained identical UL146 G12 sequences. UL139 G2, which was identified in 27% of samples, was represented by identical sequences from Hungary, the UK and The Gambia and also identical sequences in China and Germany.

3.8 Linkage between UL146 and UL139

UL146 and UL139 are 5.2 kbp apart on the HCMV genome (Figure 1.3). Potential linkage between the genotypes of these two genes was investigated in the 60 strains for which single genotypes of both UL146 and UL139 were obtained. A total of 112 genotype pairs are possible (14 UL146 genotypes X 8 UL139 genotypes). As shown in Table 3.12, 41 of these were observed at least once. Potential linkage between genotype pairs was assessed by a chi-square test. No significant variation from the expected distribution was observed, with one exception, across UL146 G9 row (*p*=0.02).

Table 3.12: Analysis of Linkage Disequilibrium

UL146 Genotype	*UL139 genotype*									*Chi-square (p)*
	G1	G2	G3	G4	G5	G6	G7	G8		
G1	1	0	0	3	1	0	0	0		0.88
G2	1	1	0	0	0	0	0	0		0.99
G3	1	0	0	0	0	0	0	0		1.00
G4	0	0	1	1	0	0	0	1		0.06
G5	0	1	0	1	0	0	1	0		0.58
G6	0	0	0	0	0	0	0	0		-
G7	1	4	1	3	2	1	0	0		0.98
G8	0	1	0	1	1	0	1	0		0.73
G9	0	1	0	0	0	2	1	0		0.02
G10	0	3	0	1	0	0	0	0		0.99
G11	0	2	0	1	0	0	0	0		1.00
G12	1	1	1	1	2	1	0	0		0.71
G13	2	2	0	5	1	0	0	1		0.76
G14	1	0	0	0	0	0	0	0		1.00
Totals	8	16	3	17	7	4	3	2	60	

3.9 Infections with multiple HCMV strains

Multiple genotypes for one or both genes (UL146 and UL139) were detected in at least 14% of samples upon first analysis. Approximately one third of all samples were tested on three separate occasions to assess reproducibility of the results. As shown in Table 3.1, a number of these samples tested in triplicate yielded additional genotypes not found in the initial experiment. In addition to this, repeat experiments occasionally failed to detect the original genotype detected. When results from repeat experiments were included, the number of mixed infections increased to 29%. Thus, in the initial experiment, more than one genotype was detected in 11% of European samples, 16% of Gambian samples, 47% of South African samples and 10% of Hong Kong samples. This increased to 24%, 33%, 60% and 60%, respectively, when repeat experiments were included. Table 3.13 displays the age and clinical details of those samples containing multiple genotypes. Multiple genotypes were distributed equally among adults and infants. The majority of samples containing multiple genotypes were from immunocompromised individuals (81%), however 90% of samples sequenced were from immunocompromised individuals. Therefore, mixed infections were identified in 21% of samples from immunocompromised samples and 42% of samples from immunocompetent individuals.

Table 3.13: Immune status and age of patients with multiple infections

Strain[a]	*Age/sex*[b]	*Details*[c]	*UL146 genotype*[d]	*UL139 genotype*[d]
A4	9/M	B	G9 (G13)	G7 (-)
A5	1/F	B	G4 (-)	G2, G7
A6	15 d/F	J/C	G13 (-)	G1, G4, G7 (-)
A11	Infant	C	ND	G4, G6
C1	Adult	KT	ND (G12)	G5 (G4)
C2	Infant	C	G2 (G9)	G6 (G3, G5)
C3	Infant	C	G6, G9 (G7)	G3 (G1, G4)
C7	Infant	C	G7 (-)	G3 (G4)
C8	Infant	C	G1 (-)	ND (G2, G3, G8)
C10	Infant	C	G1 (G7)	G1 (G8)
D1	?	?	G5, G9	ND
D2	?	?	G7 (G3, G12, G13)	ND (G1, G4)
D5	?	?	G12 (G4, G7)	ND (G2, G3, G7)
E4	Infant	?	G13 (-)	G7 (G1)
E5	3	?	G9 (G7, G12)	ND (G4)
E6	Infant/M	T	G2	G1, G2
E7	Adult/M	LT	ND (G13)	G4, G6 (-)
E10 *(AL)*	Adult	H	G10 (-)	G5 (G4)
E12 *(W)*	Adult	H	G13 (G7)	G1, G2, G7 (-)
G2	Infant	C	G13 (-)	G2, G3 (-)
G3	Infant	C	ND	G2, G7

G4	Infant	C	G13 (-)	G4 (G7)
G8	Infant	PN	G5 (G12, G13)	G4 (G1, G2)
G17	Infant	PN	G8	G1, G5, G8
G18	Infant	PN	G3 (G7, G8, G12)	G1 (-)
H3	Infant	C	G9	G4, G7
I3	Adult/F	P	G7, G13	G4
I5	4 m/M	?	G12 (-)	G4 (G5)
S2	45/F	C+	G7 (G13)	G2 (G4)
S3	10/M	B	G7, G12	G2
S4	35/M	B	G9, G12, G13 (G2, G7)	G2, G5 (G1)
S7	56/M	AP	G7 (-)	G6 (G5)
S12	2/M	W	G7 (G2, G9)	G2 (G6)
S13	0/F	P	G2 (G4)	G2 (G6)
S23	49/M	R	G10 (G12)	G5 (G4)
S26	48/F	CCO	G1, G10, G13	G5
S33	40/M	?	G13 (G9, G12)	G4 (G1, G5)
S43	Adult	KT	G13	G1, G4
S44	Adult	KT	G2	G6, G7
S45	?	KT	G1 (-)	G6 (G3)
Z1	28/F	H/C+	ND	G2, G5
Z2	36/F	C+	G1, G13 (G12)	G5 (G1, G4)
Z3	22/F	C+	G9, G13 (-)	G1 (-)
Z4	20/F	C+	G9 (G5)	G2, G5 (-)
Z5	18/F	C+	G3, G14	G2
Z6	20/F	C+	G13 (G1, G7)	ND (G4, G5)
Z7	29/F	C+	G1, G3, G9 (G7)	G4 (G5)
Z11	26/F	H/C+	G1 (G12)	G1 (G4, G5)
Z14	30/F	C+	G3, G7, G13 (G12)	G5 (G2)

[a] A, Australia; C, Hong Kong; D, Germany; E, England; G, The Gambia; H, Hungary; I, Italy; S, Scotland; Z, South Africa. The original strain designations of sequences listed in a previous study (Dolan *et al.,* 2004) are given in parentheses.

[b] Ages are in years unless specified in days (d), weeks (w) or months (m). Sexes: F, female; M, male; ?, unknown.

[c] AP, aplastic anaemia with pneumonia; B, bone marrow transplant; C, congenital; C+, HCMV positive by PCR or IgM in serum; CCO, Crohn's disease with HCMV colitis; H, HIV positive; J, jaundice; KT, kidney transplant; LT, liver transplant; P, maternal HCMV positive (pregnancy); PN, postnatal; R, renal transplant; T, thrombocytopenia; W, Wilms' tumour post-nephrectomy.

[d] Genotypes are denoted G1-G14 for UL146 and G1-G8 for UL139. Multiple genotypes are separated by commas. ND, not determined. Additional genotypes identified in subsequent experiments are in parentheses; -, no additional genotypes identified.

3.10 Computer modelling of UL146 genotypes

The high degree of variation between the UL146 genotypes could indicate functional differences between the protein products. To investigate this, homology models of each UL146 genotype were constructed using the MOE protein modelling and 3D bioinformatics software (CCG), since the 3-D structure of UL146 has not been determined.

Comparative modelling methods generally use structural templates that have the highest sequence similarity to the target protein. MOE contains a built-in library of experimentally determined high-resolution protein structures. A single template approach was used with application of the parameter 'amber 99 forcefield', which is specific for smaller molecules. Twenty-five intermediate models were generated, and the final model was taken as the Cartesian average of all the intermediate models. The stereochemical quality of the polypeptide backbone and its side chains in this final model was evaluated using Ramachandran plots. A Ramachandran plot is a way to visualise dihedral angles between aas in a protein structure. Rotation about the N-C bond of a peptide backbone is denoted by the dihedral angle φ, and rotation about the C-C bond is denoted by the dihedral angle ψ. The values of φ and ψ are constrained geometrically due to steric clashes between non-neighbouring atoms. The permitted values were originally determined by Ramachandran and are usually indicated on a two-dimensional (2-D) map of the φ-ψ plane (i.e. Ramachandran plot), (Ramachandran, 1963; Creighton, 1993). Bad dihedral angles (termed outliers) plus a single residue on either side were selected and energy minimized, as a means of reducing the number of bad dihedral angles in the final model without altering the rest of the model.

The entire aa sequence (including predicted signal peptide) of a representative of each UL146 genotype (as used previously, Figure 3.2) was input to MOE and used to search for similar proteins with known crystal structures in the MOE structural family database library. The MOE search uses a FastA-type local alignment tool for aa sequence and structure alignment of protein chains, followed by a family membership test based upon the alignment produced and Z-score significance testing. A final adjustment of the similarity scores is carried out by Bayesian-based secondary structure prediction.

Only six UL146 genotypes (G2, G5, G6, G7, G13 and G14) found matches to proteins in this library. The complete aa sequences of the remaining eight UL146 genotypes were then used to search for similar proteins in the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) using the ExPASy server. Table 3.14 displays the percentage sequence similarity between each UL146 genotype and its closest match in RCSB PDB or the MOE library.

Table 3.14: Amino acid Sequence Similarity between UL146 Genotypes and Homology Templates

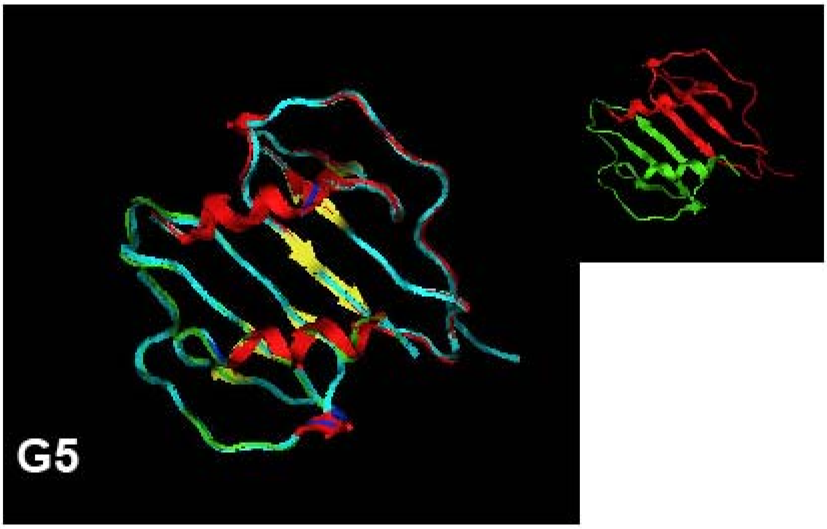
Genotype	*Similarity to best protein match in MOE[a] or RCSB PDB[b] (%)*	*Similarity to IL-8 AB (%)*
G1	24.8 (Nucleoredoxin-1) [b]	28.0
G2	35.6 (gro-α AB) [a]	36.6
G3	15.2 (IPCA, prostate cancer-associated protein) [b]	26.8
G4	23.6 (2PO7, a ferrochelatase) [b]	28.2
G5	30.3 (1F9s AB, platelet factor 4) [a]	25.4
G6	37 (1F9s AB) [a]	28.2
G7	24.2 (1F9s AB) [a]	25.4
G8	23.7 (Nucleoredoxin-1) [b]	26.8
G9	27.5 (vMIP-I AB, KSHV macrophage inflammatory protein) [b]	25.4
G10	15.6 (1XK1, human heme, oxygenase-1)	23.9
G11	30.1% (gro-α AB) [b]	23.9
G12	0 hits	25.4
G13	20.2 (1Z27, ookinete surface protein Pvs25) [a]	26.8
G14	33.3 (vMIP-I AB) [a]	28.2

[a]Most similar protein as determined by MOE
[b]Most similar protein as determined by RCSB PDB (no match found in MOE)

Initial models were generated using the template with greatest sequence similarity to each genotype as determined by MOE or RCSB PDB (Table 3.14). Figure 3.19 displays UL146 G5 modelled on the chemokine 1F9s AB and G11 modelled on the chemokine gro-α AB. Superposition of all 25 intermediate models for G5 and G11 shows good agreement and the majority of each structure remains highly conserved with the homology template, with no regions of substantial variability between intermediate models. The final energy minimized model for both molecules is also shown to the right in Figure 3.19. Table 3.14 also shows sequence similarity between each UL146 genotype and the interleukin 8 A and B chains (IL-8 AB, a functional homologue of UL146 (Chapter 1, Section 1.9), where the A and B subunits are identical).

For many of the UL146 genotypes, the level of similarity between each genotype and IL-8 is actually higher than that observed for their closest homologous protein as determined by MOE or ExPASy (with the exception of UL146 G5, G6, G9, G11 and G14). For this reason and also the observed functional homology between UL146 (Toledo) and IL-8, the aa sequence of each genotype was modelled on the solved crystal structure of IL-8 AB. Nuclear magnetic resonance and crystallographic studies indicate that IL-8 forms dimers in solution, each consisting of six stranded antiparallel β-sheets (three strands from each subunit) and two antiparallel α-helices (one from each subunit) that lie across the β-sheet (Clore *et al.,* 1990; Baldwin *et al.,* 1991). The structures of other chemokines, including gro-α, are similar to that determined for IL-8, and they also exist as dimers in solution (reviewed by Baggiolini *et al.,* 1997). However, IL-8 has more recently been demonstrated to be functionally active as both a monomer (Baggiolini *et al.,* 1995) and a dimer (Leong *et al.,* 1997).

Figure 3.20 shows superposition of all 25 models generated for UL146 G5 and G8 modelled on IL-8 AB. The majority of each structure remains highly conserved with the homology template, with no regions of substantial variability between intermediate models. The final energy models are also shown in Figure 3.20. G5 contains a single outlier, a serine residue at position 55 on the B chain. G8 contains seven outliers but, as can be seen from Figure 3.20, contains no regions of substantial variability between intermediate models. The final models for G5 and G8 are similar, both consisting of six antiparallel β-sheets and two antiparallel α-helices. Similar images were obtained for the remaining 12 UL146 genotypes when all 25 intermediate models for each were superposed (Figure 3.22). Figure 3.21 shows the final energy-minimized models obtained for these remaining 12 UL146 genotypes. All display the same conformation, six-stranded antiparallel β-sheets (three strands from each subunit) and two antiparallel α-helices (one from each subunit) that lie across the β-sheet. Figure 3.22 shows superposition of the final energy models generated for all 14 UL146 genotypes viewed from two different angles. The majority of each structure remains highly conserved, with no regions of substantial variability between each model.



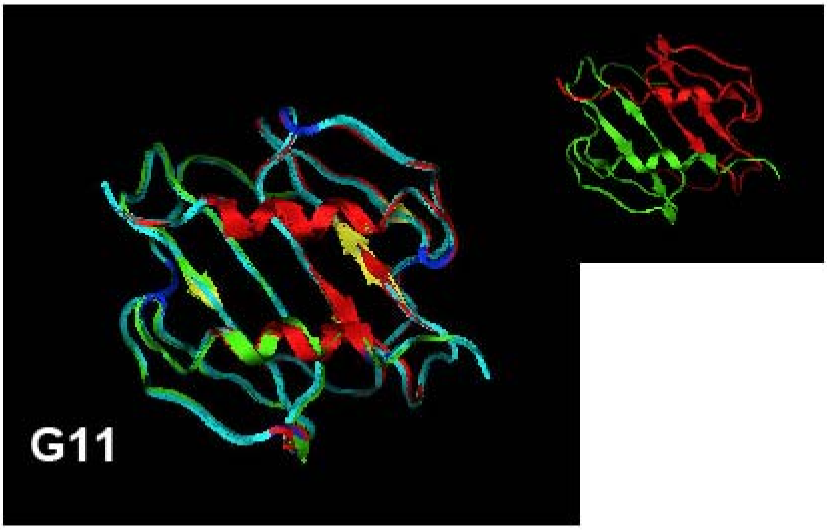


Figure 3.19 Homology models of UL146 G5 and G11 modelled on the chemokines 1F9s and gro-α

On the left is image of the superposition of 25 intermediate models (peptide backbone) for G5 modelled on 1F9s and G11 modelled on gro-α. Each intermediate model is given a different colour by MOE. On the right is the final energy minimised model (peptide backbone) for G5 and G11. The A chain is red and the B chain is green.

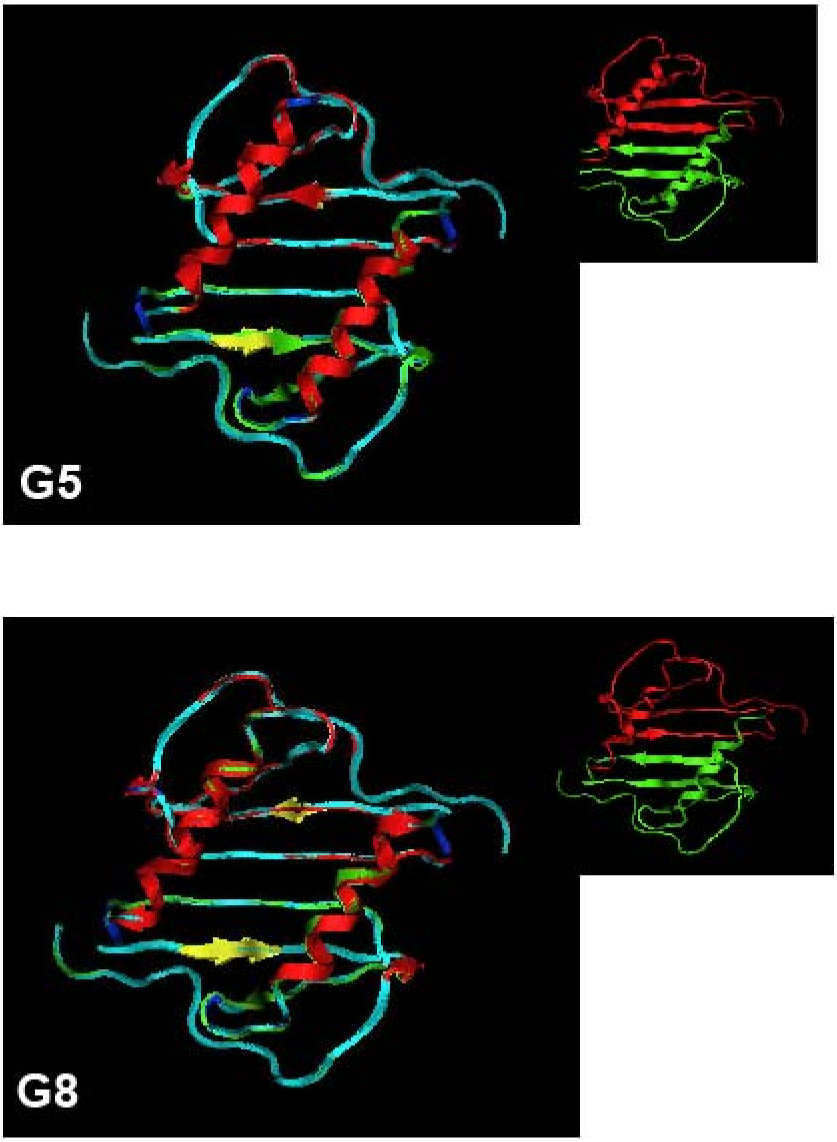


Figure 3.20 Homology models of UL146 G5 and G8 modelled on the chemokine IL-8

On the left are images of the superposition of 25 intermediate models (peptide backbone) for G5 and G8 modelled on the chemokine IL-8. Each intermediate model is given a different colour by MOE. On the right are the final energy minimised models (peptide backbone) for G5 and G8.
The A chain is red and the B chain is green

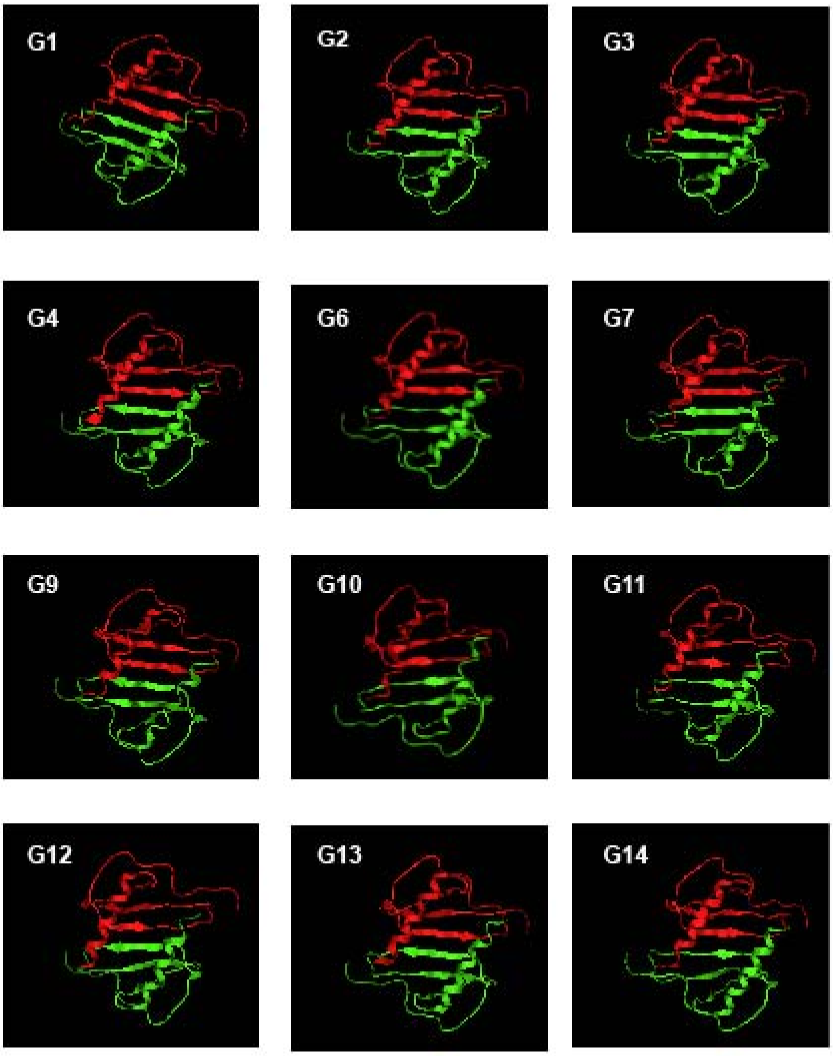


Figure 3.21 Homology models of UL146 genotypes (excluding G5 and G8) modelled on the chemokine IL-8

Final energy minimised homology models of UL146 G1-G14 (excluding G5 and G8) modelled on IL-8 AB. The A chain is red and the B chain is green.

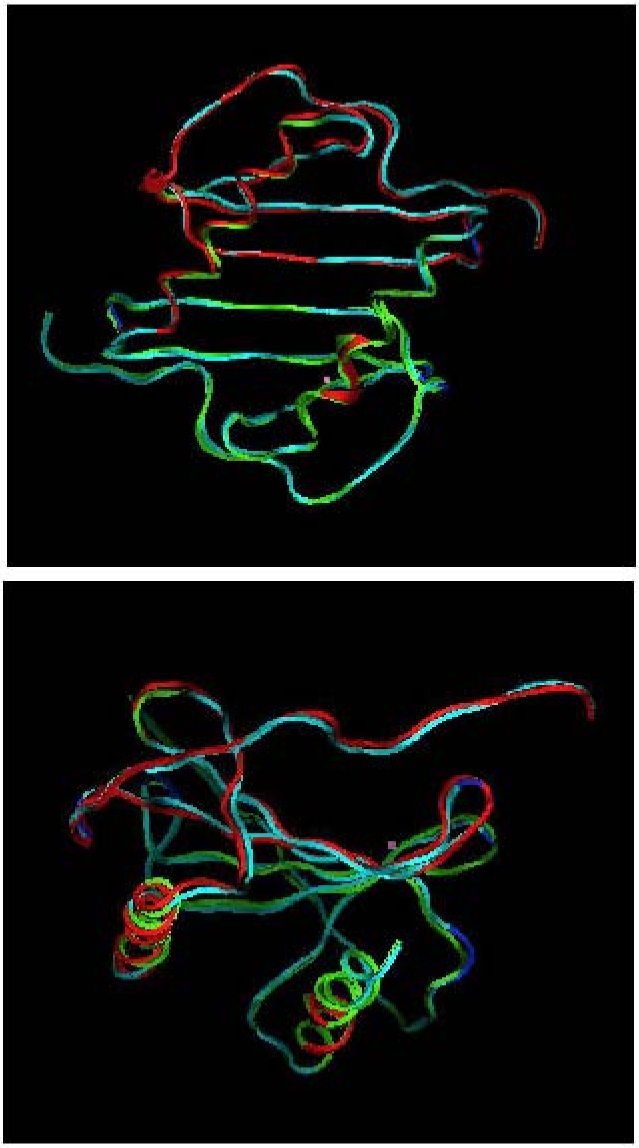


Figure 3.22 Superposition of final energy models obtained for each UL146 genotype (G1-G14) modelled on the chemokine IL-8

Final energy minimised homology models of UL146 G1-G14 modelled on IL-8, superposed on top of each other and shown from two different angles.

3.11 QPCR using UL146 genotype-specific primers

As discussed previously (Section 1.8), multiple HCMV genotypes were found in 14% of samples and this number increased to 29% when data from repeat experiments were included. This suggests that the methodology used here and in most genotyping studies could result in an underestimation of the true frequency of mixed infections. A particular sample could contain a range of concentrations of genotypes, and the outcome of PCR could vary because of stochastic effects within the reaction, resulting in a particular strain being amplified first (usually the major strain). This section describes an initial assessment of the utility of QPCR using UL146 genotype-specific primers designed to distinguish between genotypes.

To develop an assay for ascertaining the full extent of mixed infections, primers specific for each UL146 genotype were tested in a quantitative PCR assay (QPCR) utilising SYBR Green dye. SYBR Green is a fluorophore whose fluorescence increases 1000-fold in the presence of double-stranded DNA (dsDNA). It allows the measurement of amplification in real time, as the fluorescence intensity increases in proportion to the amount of PCR product. QPCR is therefore more sensitive than PCR, which relies on end point detection of amplified DNA.

Four genotype-specific primer pairs were designed to amplify UL146 G1, G2, G5 and G7 (Table 3.15). These genotypes were chosen based on the availability of template DNA confirmed to contain a single genotype. Primers were also designed for the other genotypes, but initial QPCR experiments were not followed up because of uncertainty about the templates representing a single genotype. The primers were chosen in regions that are conserved between all strains within each genotype so that they would produce amplicons of different sizes (113, 106, 132 and 143 bp, respectively). The primer design ensured that all primers had predicted melting temperatures of 60-65°C, and were predicted not to bind to other UL146 genotypes. Primers were checked to ensure they had no secondary structure using the Sigma-Aldrich website (www.sigmaaldrich.com), since primer dimers can result in artifactual fluorescence.

The presence of a particular genotype in a sample subjected to QPCR is deduced from fluorescence above a threshold and from the presence of a specific peak in the melting (dissociation) curve for the product. The baseline fluorescence is established between PCR cycles 3-15, and the threshold can then be set (automatically or manually), and Ct (i.e. the cycle threshold) is defined as the cycle at which fluorescence reaches the threshold level.

Table 3.15: Genotype-specific UL146 Primers for QPCR

Genotype	*Primer*	*Sequence (5'-3')*	*PCR product size (bp)*
G1	UL146G1FWD	GCATATGTGTATCATTATGAGGTG	
	UL146G1REV	GGATCGCGGATGAAGCCAATA	113
G2	UL146G2FWD	GGAATTACGCTGCAAATGTC	
	UL146G2REV	GTTATTGCATCTGGGACCACC	106
G5	UL146G5FWD	CTGAAGGTAATGGTCGTTGT	
	UL146G5REV	CTATCTTTATCATGACTTGTCCC	132
G7	UL146G7FWD	AGAGAATTGCGTTGTCCGT	
	UL146G7REV	CATACAGGTTTACCTCGAGG	143

The amplification data are displayed as a plot of Rn against the cycle number. Rn is the normalised reporter signal, which is the cycle-by-cycle ratio of fluorescence of the reporter dye, in this case SYBR Green, to that of the passive reference dye (ROX). ROX is a dye molecule that is included in the SYBR Green PCR master mix and does not interfere with the PCR.

In dissociation analysis, all PCR products were melted at 95°C, annealed at 55°C, and subjected to a gradual increase in temperature to 95°C. Fluorescence data (in standard units) were collected during incremental increases in temperature. A characteristic dissociation curve was obtained by plotting fluorescence against increasing temperature. Due to differences in amplicon size and nucleotide composition of each amplicon, a different dissociation curve should be obtained for each genotype. The genotype-specific primers were first tested (in triplicate) on the appropriate templates to confirm specificity. A template for each of these genotypes was produced by amplifying UL146 from a sample known to contain a single HCMV strain, using primers in conserved regions outside UL146 ORF (Chapter 2, Table 2.2.). The PCR products were quantified using a spectrophotometer.

Table 3.16: QPCR Template DNAs for Testing Genotypic Primers

UL146 Genotype	*HCMV Strain*
G1	U4 *(Toledo)*
G2	W9 *(Merlin)*
G5	U3 *(Davis)*
G7	U5 (*Towne*)

A schematic diagram of one half of the 96-well plate layout for QPCR is shown in Figure 3.23. A specific genotype template was added to all wells with the exception of the no template control (NTC). Two genotypes were tested on each plate. Primers specific for HCMV UL54, which is a highly conserved gene encoding DNA polymerase, were designed and are shown in Table 2.2 (Chapter 2). Three wells containing the template being tested and UL54-specific primers were included as an endogenous control. The endogenous control is included in order to normalise the results, as there may be variations in the amount of input DNA from well to well.

Figure 3.24 shows the results from the specificity test of the four pairs of genotype-specific primers with G5 template DNA. For the control wells A1-A3, no amplification was detected. Wells A4-D12 contained G1-G14 genotypic primers (in triplicate) and UL146 G5 template. Amplification was observed only in wells B4-B6, which contained G5 primers and G5 template. No amplification was observed in wells D10-D12 that contained primers specific for UL54. The starting level of fluorescence was higher in wells containing UL54 primers and it remained relatively constant throughout all cycles. As there appeared to be insufficient genomic DNA present in the samples for amplification of UL54, the endogenous control was replaced with a positive control (G1 template with G1 primers).

The results from the specificity test of all genotype-specific primers with G1, G2 and G7 template DNA are shown in Figures 3.25, 3.26 and 3.27, respectively.

	1	2	3	4	5	6	7	8	9	10	11	12
A	NTC + G1 PRIMERS	NTC + G1 PRIMERS	NTC + G1 PRIMERS	GA + G1 PRIMERS	GA + G1 PRIMERS	GA + G1 PRIMERS	GA + G2 PRIMERS	GA + G2 PRIMERS	GA + G2 PRIMERS	GA + G3 PRIMERS	GA + G3 PRIMERS	GA + G3 PRIMERS
B	GA + G4 PRIMERS	GA + G4 PRIMERS	GA + G4 PRIMERS	GA + G5 PRIMERS	GA + G5 PRIMERS	GA + G5 PRIMERS	GA + G6 PRIMERS	GA + G6 PRIMERS	GA + G6 PRIMERS	GA + G7 PRIMERS	GA + G7 PRIMERS	GA + G7 PRIMERS
C	GA + G8 PRIMERS	GA + G8 PRIMERS	GA + G8 PRIMERS	GA + G9 PRIMERS	GA + G9 PRIMERS	GA + G9 PRIMERS	GA + G10 PRIMERS	GA + G10 PRIMERS	GA + G10 PRIMERS	GA + G11 PRIMERS	GA + G11 PRIMERS	GA + G11 PRIMERS
D	GA + G12 PRIMERS	GA + G12 PRIMERS	GA + G12 PRIMERS	GA + G13 PRIMERS	GA + G13 PRIMERS	GA + G13 PRIMERS	GA + G14 PRIMERS	GA + G14 PRIMERS	GA + G14 PRIMERS	GA + UL54 PRIMERS	GA + UL54 PRIMERS	GA + UL54 PRIMERS

Figure 3.23 Schematic diagram of QPCR plate layout

Wells A1-A3 (highlighted by red boundaries) contain G1 primers but no template (no template control (NTC)), and are negative controls. Wells A4-D9 contain genotype A (GA) template DNA and each genotypic primer pair (in triplicate). The genotype-specific primer name is colour highlighted to show triplicate wells. Wells D10-D12 contain GA template DNA and UL54-specific primers and are endogenous controls. The diagram shows only half of a 96-well plate.

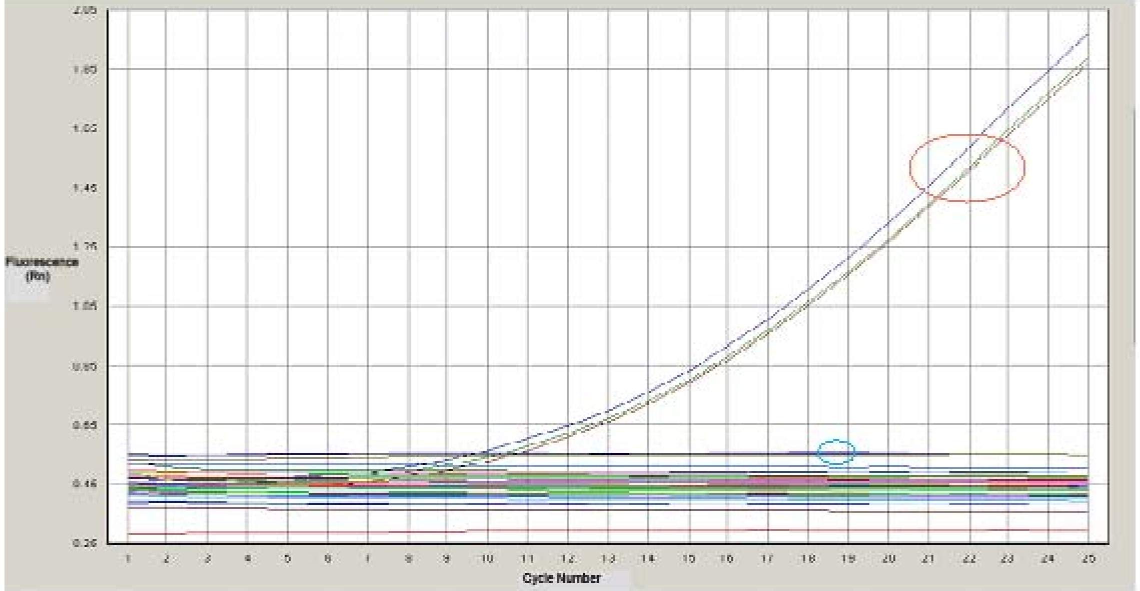


Figure 3.24 QPCR amplification of UL146 G5 template using genotype-specific primers

See Figure 3.23 for the plate layout. The ABI program arbitrarily applies a different colour to each well, hence the differently coloured lines. Amplification was observed for the three wells (B4-B6) containing UL146 G5 template DNA and G5 primers, as indicated by the red circle. Wells D10-D12 contained UL54-specific primers and are indicated by a blue circle.

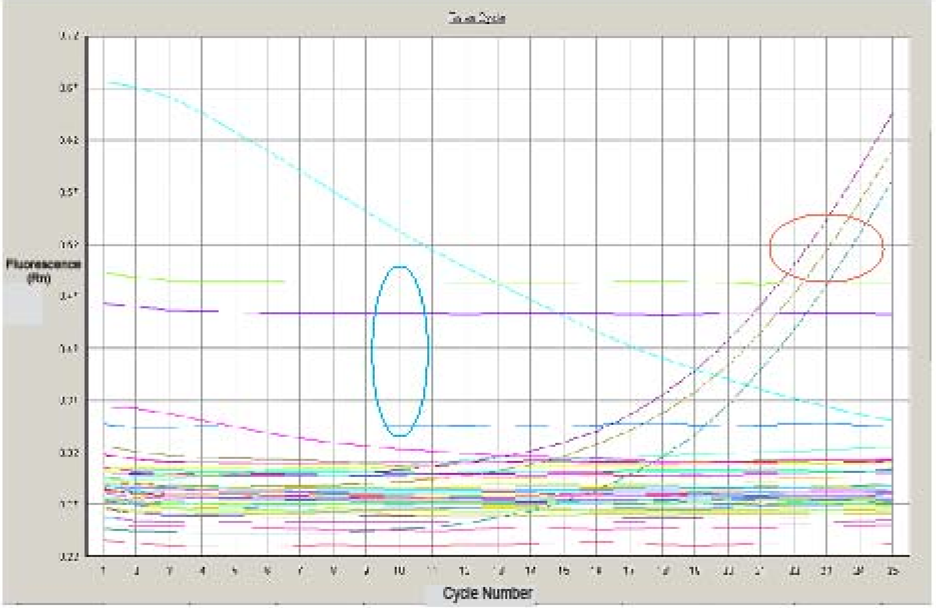


Figure 3.25 QPCR amplification of UL146 G1 template using genotype-specific primers

See Figure 3.23 for the plate layout. The ABI program arbitrarily applies a different colour to each well, hence the differently coloured lines. Amplification was observed for the three wells (A4-A6) containing UL146 G1 template DNA and G1 primers, as indicated by the red circle. Wells D10-D12 contained UL54-specific primers and are indicated by a blue circle. The blue decending line contained G11-specific primers.

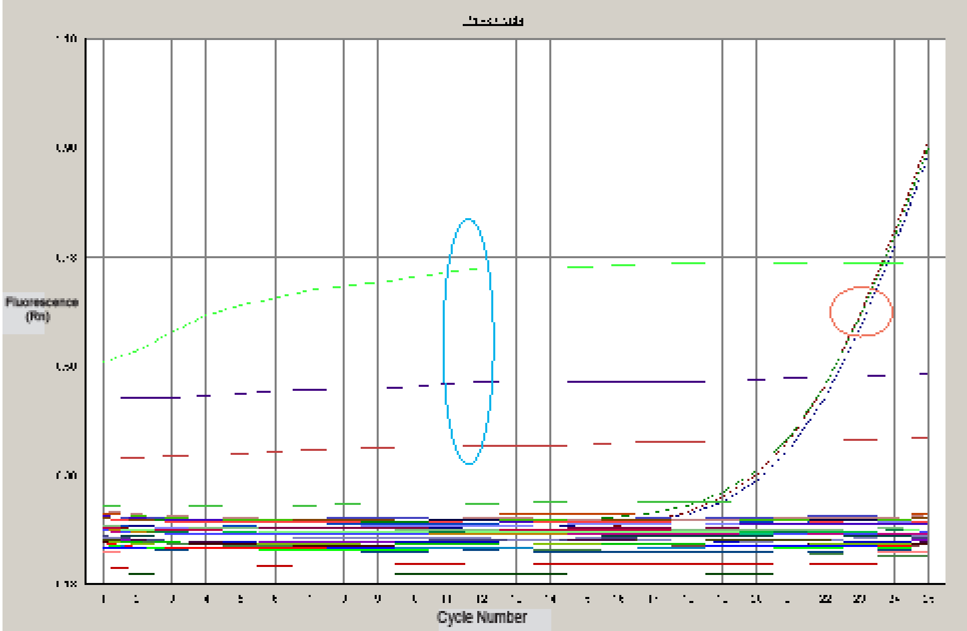


Figure 3.26 QPCR amplification of UL146 G2 template using genotype-specific primers

See Figure 3.23 for the plate layout. The ABI program arbitrarily applies a different colour to each well, hence the differently coloured lines. Amplification was observed for the three wells (A7-A9) containing UL146 G2 template DNA and G2 primers as indicated by the red circle. Wells D10-D12 contained UL54-specific primers and are indicated by a blue circle.

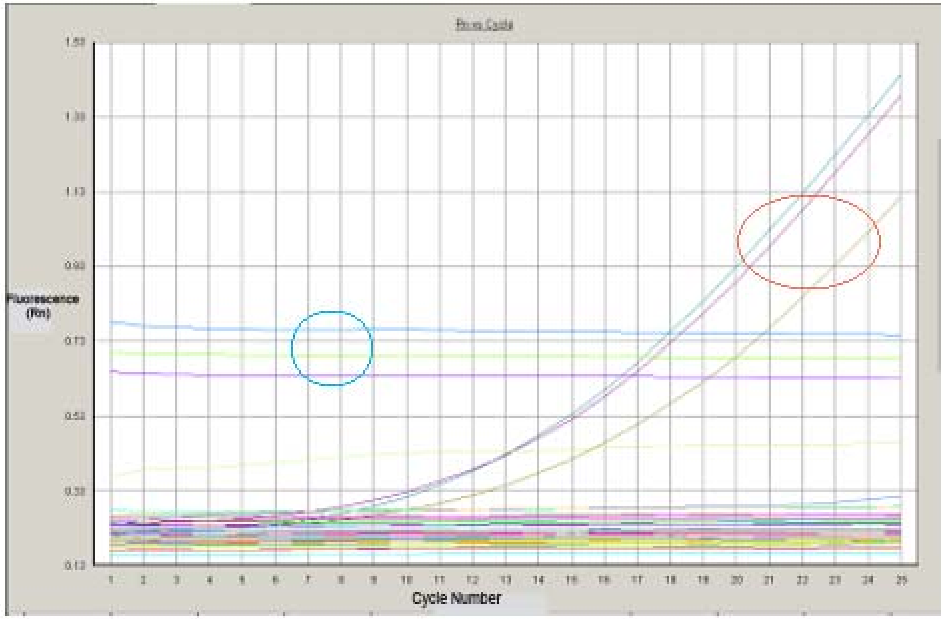


Figure 3.27 QPCR amplification of UL146 G7 template using genotype-specific primers

See Figure 3.23 for the plate layout. The ABI program arbitrarily applies a different colour to each well, hence the differently coloured lines. Amplification was observed for the three wells (B10-B12) containing UL146 G7 template DNA and G7 primers, as indicated by the red circle. Wells D10-D12 contained UL54-primers and are indicated by a blue circle.

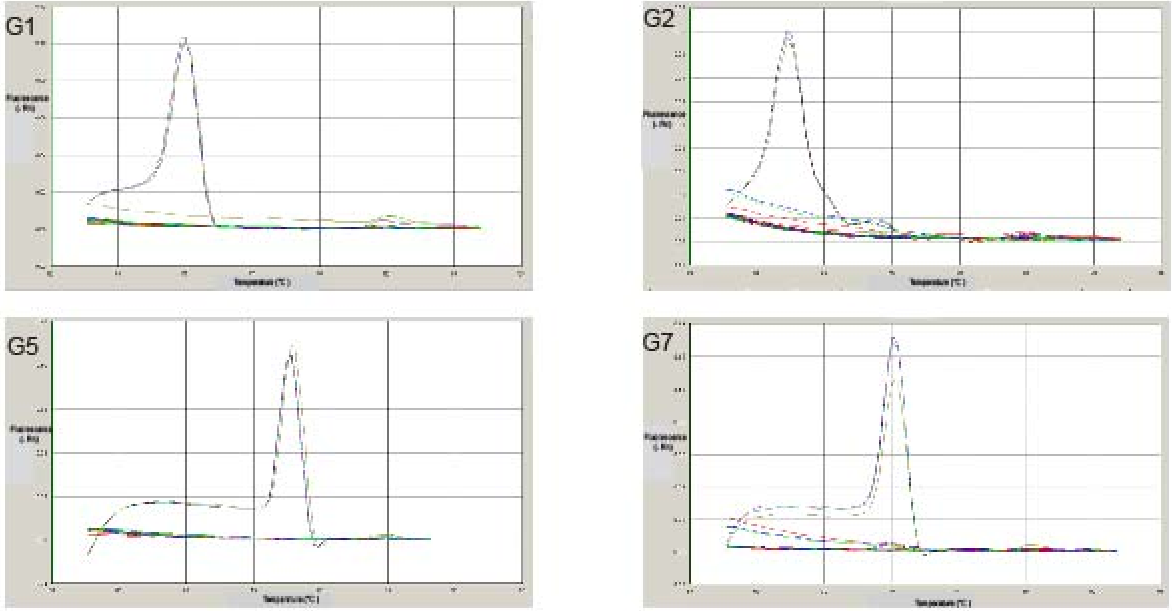


Figure 3.28 Dissociation curves for UL146 G1, G2, G5 and G7 QPCR products

A plot of the negative first derivative of raw fluorescence data versus increasing temperature measured during the melting curve of all QPCR products produced on each half plate for testing of G1, G2, G5 and G7 template DNA with all UL146 genotype-specific primer pairs. The ABI program arbitrarily applies a different colour to each well.

As observed for G5, no amplification was detected in the negative control wells and amplification was observed only in the wells containing genotype-specific primers that matched the template DNA. For those wells containing UL54-specific primers (D10-D12), again no amplification was observed, although the starting level of fluorescence was even greater than that observed for the G5 template DNA test and remained relatively constant throughout all cycles.

Specific melting curve peaks were obtained for all the genotypes tested, but there is the potential for overlap between some genotypes, which means this measurement alone is insufficient for genotype detection. Figure 3.28 shows dissociation curves obtained for wells containing UL146 G1 genotype specific primers and G1 template (Tm of ~70°C), G2 genotype specific primers and G2 template (Tm of ~ 67°C), G5 genotype specific primers and G5 template (Tm of ~ 78°C), and G7 genotype specific primers and G7 template (Tm of ~ 75°C). No specific melting curves were obtained for any of the other wells on the plate, including those wells containing UL54 primers.

The specificity of genotype specific primers was confirmed by agarose gel electrophoresis of QPCR products obtained. Bands were only visible for those samples that tested positive by QPCR (data not shown).

3.12 Discussion

The overall aim of this study was to characterize UL146 and UL139 sequences in a large panel of clinical isolates collected from Africa (South Africa and The Gambia), Asia (Hong Kong), Australia and Europe (various countries). Specific aims were to determine the total number of circulating genotypes, the frequencies at which they occur and any bias in their geographical distribution, and to investigate the mode of evolution that resulted in these genotypes and whether there is any evidence for linkage between UL146 and UL139. A total of 182 UL146 sequences and 183 UL139 sequences were derived experimentally. These sequences were supplemented with 168 previously published UL146 and 117 UL139 sequences for genotyping purposes and for investigation of the mode of evolution. All UL146 sequences analysed fell into the 14 genotypes described previously (Dolan *et al.*, 2004), and no new genotypes were detected. UL146 is highly variable throughout its length, with only a few residues conserved in all sequences (Figure 3.2). Twelve genotypes contain the ELRCXC motif, which has been shown to be essential for receptor binding and IL-8 activity (Clark-Lewis *et al.,* 1991, 1995), and two contain the NGRCXC motif (G5 and G6), which has been shown to be important for interaction with T and B cells (Baggiolini *et al.,* 1997). UL146 G5 and G6 were found in only 5% of all samples, whereas genotypes with the ELRCXC motif make up the remaining 95%. This suggests that the ELRCXC motif is important in UL146 function and that the NGRCXC genotypes provide niche functions in specific situations. UL146 G4 appears to be most closely related to the CCMV UL146 sequence. This could indicate that UL146 G4 is closer to the ancestral UL146 sequence, and that all other genotypes have diverged to a greater extent (Figure 3.3). However, the tree in Figure 3.3 was produced using a single CCMV UL146 sequence and has no molecular clock. It is likely that CCMV UL146 sequences have also undergone divergence and that the single CCMV UL146 sequence available is not representative of all those sequences in circulation in chimpanzees. Sequencing of additional CCMV isolates is required to investigate this question further.

All UL139 sequences grouped into eight genotypes. A recent analysis of 26 clinical samples (Qi *et al.,* 2006) described three major groups (G1, G2 and G3), two of which were divided into subgroups (G1 into G1a, G1b and G1c and G2 into G2a and G2b). Subgroups G1b and G1c in the previous study correspond to G1 in the present study, subgroup G1a corresponds to G4, subgroups G2a and G2b correspond to G6 and G2, respectively, and G3 is named identically in both studies. Unlike UL146, variation in UL139 is concentrated in a region towards the N terminus and is due to substitutions or deletions of variable size in a region that is rich in S and T residues, and likely to contain *O*-linked glycosylation sites. This region also contains NXS or NXT motifs that may be *N*-linked glycosylation sites. The number of possible *O*-linked and *N*-linked glycosylation sites varies between UL139 genotypes, and it may be that selection focuses primarily on the glycosyl side chains rather than the underlying aa sequence. This feature has been described in other variable glycoprotein genes, such as UL73 (gN) and UL74 (gO) (Mattick *et al.,* 2004; Pignatelli *et al.,* 2001, 2002, 2003).

The analysis suggests that constraint has been the predominant factor in the evolution within UL146 and UL139 genotypes, with positive selection detected marginally at best. Although the sequence alignments within each UL146 and

UL139 genotype appeared reliable, the high level of variation made the alignments unreliable, particularly in the most variable regions. Investigation of the mode of selection that may have resulted in the diversification of the genotypes depended on the quality of the alignments, and thus the conclusions drawn within genotypes, rather than among them, are more reliable. For UL146 the analysis suggests that most genotypes have evolved under constraint or neutrality. Some evidence for positive selection was found for UL146 G1, G2, G3, G4, G9, G10 and G13. However, further analysis using *codeml,* which tests different codon-substitution models, suggests that this is not significant for G2, G3, G4, G9, G10 and G13 and significant only at the 5% level for G1. Investigation of the mode of selection among UL146 genotypes also suggests they have evolved under constraint rather than positive selection. This is in agreement with another study that investigated the mode of selection among 25 UL146 sequences (Arav-Boger *et al.,* 2005). For UL139, all the analyses suggest that within and among genotypes the genes have evolved mainly under constraint or neutrality. This was concluded from analyses performed using a number of programmes, including pairwise dN/dS in Swaap and the codon based Z-test in Mega 4.0.

Within and among genotypes, both genes appear to be under constraint rather than positive selection. It could be that the genotypes diverged early in human history and nonsynonymous changes became fixed as a result of purifying selection. Overall, the most likely scenario is that the genotypes developed in early human populations (or even earlier), becoming established via founder or bottleneck effects, and have spread and mixed worldwide in more recent times. The founder effect is the loss of genetic variation that occurs when a new colony is established by a small number of individuals from a larger population (Hey, 2005). Originally, different genotypes may well have shown distinct geographical locations but are now found in all regions.

In general, there is little evidence for linkage between UL146 and UL139 genotypes. This is in accordance with other studies investigating potential linkage between variable HCMV genes (Rasmussen *et al.,* 2003). However, it should be noted that the analysis of UL146 and UL139 might have been compromised by the relatively small sample number (60) in relation to the large number of possible genotype combinations (112). Chi-square analysis of UL146

G9 revealed a departure from the expected distribution (*p*=0.02), as there were two occurrences of UL146 G9 with UL139 G6. However this is likely a consequence of the fact that both samples were isolated from patients in the same location. In general, genetic linkages between HCMV genes are rare, and have only been found between genes that are proximate on the genome or share functional interactions, such as gN and gO (Mattick *et al.,* 2004).

In general, the occurrence of UL146 and UL139 genotypes was independent of geographical source. This is supported by the similar pattern of distribution observed for both UL146 and UL139 genotypes in European and African samples (Figure 3.17 and Figure 3.18). However, some bias in the geographical distribution of UL146 genotypes was evident from the analysis. This may reflect low sample number (albeit much larger than those utilized in previous studies) and the lack of detailed information on the ethnic origin of the samples. The apparent geographic isolation of some genotypes (UL146 G10 and G11 in particular) in European samples (Table 3.8, Figure 3.17) may reflect the availability of more samples from Europe than other regions. A number of sequences available in Genbank also fall into UL146 G10 and G11, some of which were from Chinese samples, which suggests that these genotypes are indeed found outside Europe. Similarly, the single example of UL146 G6 was detected in an Asian sample (*p*=0.006, Table 3.8). This is most likely due to the rarity of this genotype, as only two UL146 G6 sequences have been detected (in this work and Dolan *et al.,* 2004).

Similarly, some bias in the geographical distribution of UL139 genotypes was detected, for UL139 G7 (*p*=0.006) in particular. However, this finding may have been compromised by insufficient sample numbers in some regions and a lack of information on the ethnic origin of the samples (Table 3.9). Indeed, UL139 G7 sequences from all regions are available in Genbank. Thus overall it appears that for UL146 and UL139, all genotypes are found in all regions, which is in agreement with another large-scale genotyping study that investigated the geographical distribution of UL73 (gN) and UL74 (gO) genotypes (Pignatelli *et al.,* 2003).

In general, there is no convincing evidence for association between genotype and pathogenicity. All UL146 and UL139 genotypes were distributed among clinical samples from both immunocompetent and immunocompromised individuals, and all were represented in urine samples. Several studies have been published on whether particular genotypes for variable HCMV genes (such as UL55, UL73, UL74 and UL146) are associated with disease outcome (Dal Monte *et al.,* 2004; He *et al.,* 2006; Pignatelli *et al.,* 2003). These studies have been limited by factors such as the choice of gene, the origin of samples, the number of samples, the general absence of genetic linkage between genes and the common occurrence (and probable underestimation) of mixed infections (discussed below). The results are confusing and sometimes contradictory (Arav-Boger *et al.,* 2002, 2006a; Aquino and Figuerido, 2000; Dal Monte *et al.,* 2004; Mattick *et al.,* 2004; Puchhammer-Stockl et al., 2006). Nonetheless, most are in accordance with the findings of the present study. Thus, within the limits of the sample number, no convincing broad association of genotype with disease has emerged.

Similarly, no correlation between genotype and sample type (as a potential indicator of cellular tropism or compartmentalisation) emerged. Compartmentalisation has been described for AIDS patients in respect of gB genotypes (Tarrago *et al.,* 2003), and more recently during the investigation of gB genotypes in lung and blood compartments of transplant recipients.

The Toledo-encoded UL146 protein has been shown to share functional homology with human IL-8, being capable of neutrophil degranulation, calcium mobilization and chemotaxis (Penfold *et al.,* 1999). Variation has been observed in the promoter region of human IL-8 rather than within the coding sequence as seen for UL146, and IL-8 shows differential expression in response to various stimuli and in different tissue types (Baggiolini *et al.,* 1995a, 2000). It is not known whether the UL146 genotypes possess different biological properties, and functional studies to investigate this question are required. However the function of a protein can sometimes be inferred from sequence or structural similarity. All UL146 variants contain the CXC motif and two additional cysteines found in many human chemokines. Comparative modeling is based on the general observation that evolutionarily related sequences tend to have similar 3-D structures (Chothia and Lesk, 1986).

Since Toledo-encoded UL146 has functional homology with IL-8 and IL-8 is a CXC chemokine, IL-8 was chosen as a template for homology modeling of UL146 variants (although models were also produced using other templates, such as gro-α and 1F9s AB). In solution, IL-8 exists as a dimer, of two identical subunits, each of Mr ~8000 (Clore *et al.,* 1990). Therefore, two copies of each UL146 genotype (complete aa sequence including predicted signal peptide) were homology modeled on the A and B chains of IL-8. All final energy models and, indeed, all intermediate models for all UL146 genotypes showed good agreement with the solved structure for IL-8. Despite extensive sequence variation among genotypes, each conformed to a similar 3-D structure containing an α-helix and three β-sheets (Figure 3.22). Many chemokines have similar structures. For example, human chemokines such as gro-α and 1F9s share this α-helix and three β-sheet conformation, despite having highly differing aa sequences, and they bind to the same cell surface receptors and have similar functions, albeit sometimes with differing affinities (Baggiolini *et al.,* 1997; Mayo *et al.,* 1995; Zhang *et al.,* 2000). This could indicate that, despite the high level of sequence variation, the genotypes are functionally similar.

As an initial experiment, UL146 G5 and G11 were modelled on the solved 3-D structures for chemokines 1F9s and gro-α, respectively. Good agreement was obtained between intermediate models and the template, and small numbers of bad dihedrals were derived in the final energy-minimised model (Figure 3.19). Although structurally very similar to IL-8, gro-α and 1F9s differ from IL-8 in N-terminal region, which contains the ELR motif and is thought to influence receptor binding. The receptor CXCR1 has high affinity for IL-8 and low affinity for other chemokines, whereas CXCR2 has high affinity for IL-8 as well as other CXC chemokines such as gro-α. This suggests that even small differences in the 3-D structure of UL146 genotypes might affect receptor specificity (Baggiolini *et al.,* 1997). Functional studies using different UL146 variants are required to investigate whether there are any functional differences between genotypes.

Homology modeling of each UL146 genotype (complete aa sequence) on multiple templates (Table 3.14) was performed using the programme MOE (CCG). The best structural models, those which displayed a high level of agreement between intermediate models (Figure 3.22) and small numbers of bad dihedrals in the final energy minimised model, were obtained when chemokines with an α-helix and three β-sheet conformation were used as the template (Figures 3.19-3.22). Despite the high level of aa and nucleotide sequence divergence, all UL146 genotypes have similar predicted 3-D structures (α-helix and three β-sheet conformation). This indicates that different genotypes, as with different chemokines, are functionally similar, though small structural differences may prove advantageous depending on the cell type or host encountered.

A region of sequence identity (SETTTGTSSNSSQST in Figure 3.8) has been noted between the UL139 protein and CD24, a cellular glycosyl phosphatidylinositol-linked glycoprotein that is involved in B cell activation (Qi *et al.,* 2006). This sequence is present in all of the UL139 genotypes identified in the present study, except for G5, and is also found in CCMV UL139. It is difficult to assess the significance of this similarity, especially as it is absent from UL139 G5 and is not conserved in CD24 orthologues from other mammals. However, as with UL139, variation in glycosylation has been observed in CD24, and this has been linked to differences in cell and tissue specificity (Goris *et al.,* 2006; Poncet *et al.,* 1996). Additional roles for CD24 in apoptosis and cell adhesion have been suggested, and more recently also in regulating the responsiveness of a chemokine receptor, CXCR4 (Smith *et al.,* 2006; Schabath *et al.,* 2006). The possibility that UL139 may be a CD24 homologue remains intriguing, but unproven. Preliminary studies towards characterising the UL139 protein and potential differences between genotypes are described in Chapter 4, Sections 4.4, 4.5 and 4.6.

An interesting observation was that CCMV UL139 is much larger than HCMV UL139 and contains the coding regions of separate homologues in other CMVs. The C-terminal region of CCMV UL139 is homologous to HCMV UL139, whereas the N-terminal region is homologous to rh174, an RhCMV gene that lacks a homologue in HCMV. This suggests that an ancestor of CCMV may have originally contained counterparts of both RhCMV rh174 and CCMV UL139, and that an in-frame deletion resulted in fused coding regions (effectively yielding rh174-UL139).

Mixed HCMV infections were more frequently detected in samples from certain regions, namely Hong Kong, South Africa and, to a lesser extent, The Gambia. It is possible that this is a result of higher transmission frequencies. In one study

(Beyari *et al.,* 2005), a higher seroprevalence frequency in children in Malawi compared to European countries and the USA was taken as possibly reflecting greater opportunities for transmission, but multiple genotypes were detected in only a small number of samples. In the present study, a single UL139 genotype and multiple UL146 genotypes, or vice versa, were detected in some samples. This could be due to different strains happening to contain the same genotype at one locus but not at the other, or to the limitations of amplifying different sequences present in different amounts in mixtures. A proportion of samples tested more than once were found to contain additional genotypes not apparent from the initial experiment, suggesting that the frequency of mixed infections was underestimated by the methodology used. In addition, the original genotype detected was not always detected in subsequent experiments. This is one of the limitations of PCR-based genotyping studies based on the use of conserved primers. Moreover, there is no guarantee that all genotypes will be detected, as the conserved primers are chosen on the basis of alignments of available sequences. Another potentially complicating factor is the possibility of cross contamination during DNA extraction from a clinical sample. Although all precautions were taken to avoid this, the possibility that multiple genotypes in a particular sample are a result of cross contamination cannot be ruled out with certainty.

Although the present study detected a large number of sequences for both UL146 and UL139 and all previously published sequences fell into the genotypes defined from them, there is still a possibility that some UL146 or UL139 genotypes have escaped recognition. These may be detected in future studies utilizing different or redundant primers or from whole genome sequencing. As reported in other studies, mixed infections were found in both immunocompromised (transplant recipients, neonates and AIDS) patients and healthy individuals (Table 3.1 and Table 3.14) (Arav-Boger *et al.,* 2002, 2002a, 2005, 2006; Coaquette *et al.,* 2004; Gerna *et al.,* 1992; Meyer-König *et al.,* 1998, 1998a; Puchhammer-Stöckl *et al.,* 2006). In addition, mixtures were equally distributed among infants and adults suggesting age is not a factor. Multiple infections have important implications, not just for investigation of potential association between HCMV strain and disease but also for vaccine design. Indeed, numerous virus strains with distinct immunogens complicate vaccine candidate choice. In addition, there has been some suggestion that mixed infections may have a negative impact on the course of HCMV infection in solid organ transplant recipients due to delayed viral clearance (Humar *et al.,* 2003; Puchhammer-Stöckl *et al.,* 2006).

Further work is needed to determine the true frequencies of mixed infections in clinical samples. A QPCR assay using genotypic primers and SYBR green showed promise in investigating whether the proportion of mixed infections is indeed underestimated. This assay is in the preliminary stages of development and further work is needed to validate it before it can be utilised on clinical samples. Other methods could also be employed towards this end, including QPCR using MGB Taqman probes, which are designed to be specific for a particular genotype and cannor fall foul of false positives due to primer dimers and/or non-specific PCR products. Alternatively, PCR could be carried out in the presence of an oligonucleotide substituted with locked nucleic acids. This suppresses the amplification of a specific sequence by a factor of approximately 1000, thereby enabling the amplification of a second, different sequence (Prepens *et al.,* 2007).

4 Transcript mapping of UL146 and UL139 and initial characterisation of the UL139 protein

4.1 Introduction

UL146 encodes a CXC (α) chemokine that may promote dissemination of the virus throughout the host through its ability to attract monocytes to the initial site of infection. UL139 is predicted to encode a type I membrane glycoprotein with a small region of homology to CD24, which suggests a possible immunomodulatory role for its protein product. Both genes are hypervariable and a number of studies have been published investigating UL146 sequence variation between clinical samples (Arav-Boger *et al.,* 2005, 2006; Dolan *et al.,* 2004; Hassan-Walker *et al.,* 2004; He *et al.,* 2006; Lurain *et al.,* 2006; Penfold *et al.,* 1999; Prichard *et al.,* 2001; Stanton *et al.,* 2005), and UL139 sequence variation (Qi *et al.,* 2006). The present study investigated UL146 and UL139 sequences in a large panel of clinical isolates from a number of locations worldwide and found all UL146 sequences fell into the 14 genotypes previously defined (Dolan *et al.,* 2004) and all UL139 sequences fell into eight genotypes (Chapter 3).

This chapter reports on the investigation of the transcriptional patterns of UL146 and UL139 in the HCMV strain Merlin by northern blot analysis. The 5'- and 3'-ends of mRNAs were mapped using RACE. As an introduction, Figure 4.1 shows the arrangement of ORFs at the right end of U_L from UL139 to UL132.

During the course of this work Lurain *et al.* (2006) analysed transcriptional expression of UL146 in ten clinical isolates, four by northern blot analysis and seven by RT-PCR. It was determined that UL146 is expressed with early-late (E-L) kinetics. This is in agreement with another study, which utilised microarrays to analyse transcription from Towne and found that UL146 (designated UL152 in Towne) displays E-L kinetics (Chambers *et al.,* 1999). These results differs somewhat from a previously published study that found that the UL146 protein product is expressed in Toledo at late (L) times post-infection (Penfold *et al.,* 1999). No information has been published regarding transcription of UL139.

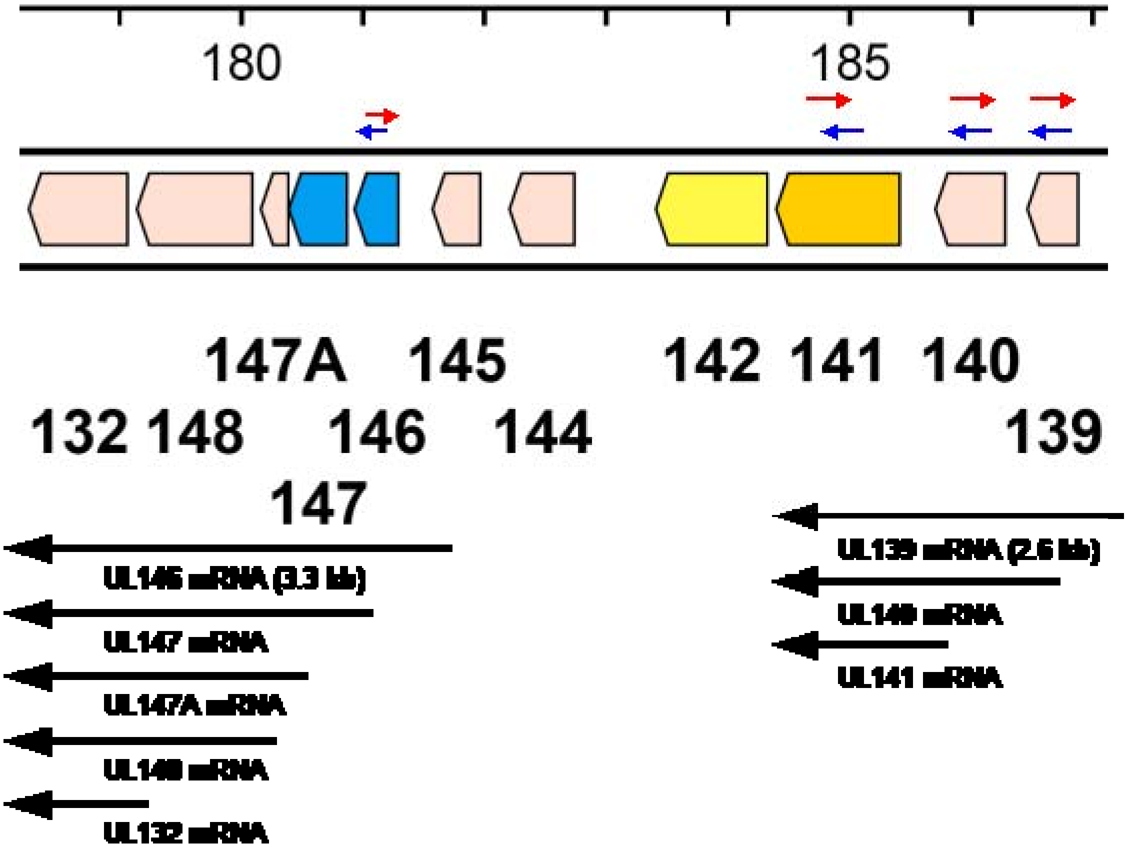


Figure 4.1 Map of HCMV ORFs at the right end of U$_L$ (UL132 to UL139)

ORFs coloured pink are non-core genes. UL142 (yellow) is a member of the UL18 family. UL141 (orange) is a member of the UL14 gene family. UL147 and UL146 (blue) are members of the UL146 (CXC) chemokine gene family. Locations in the Merlin genome are indicated by the coordinates 180 and 185 kbp (see Figure 1.3). The locations of the 5'-RACE Gene specific primers (GSPs) are indicated by red arrows and the 3'-RACE GSPs by blue arrows. Putative transcripts are indicated by leftward oriented arrows.

4.2 Northern blot experiments

Northern blot experiments were performed to determine the expression kinetics and size of transcripts from UL146 and UL139. DIG-labelled ssRNA probes were generated by amplifying the entire UL146 and UL139 protein-coding regions using the primers listed in Table 2.2. The PCR products were electrophoresed, purified, cloned into pGem-T and sequenced. Plasmids with the sequence inserted in the required orientation were selected and linearised. The DIG northern starter kit was used for *in vitro* transcriptional labelling of the probes (Chapter 2, Section 2.23.3). Duplicate blots were hybridised with a control DNA probe ($GAPDH_2$). The results are shown in Figure 4.2.

A UL146 transcript was detected in L RNA. A UL139 transcript was detected in L RNA, and to a much lower extent, E RNA. Therefore UL146 was expressed with L kinetics and UL139 was expressed with E-L kinetics. The UL146 probe hybridised to a major band approximately 3.4 kb in size and a very minor band 5 kb in size. The UL139 probe hybridised to a band approximately 2.6 kb in size.

4.3 RACE experiments

In order to map the 5’- and 3’-ends of the UL146 and UL139 transcripts, experiments were performed using the SMART RACE kit (Chapter 2, Section 2.24). This involves a PCR-based technique that amplifies the 5’- and 3’-ends of mRNAs using gene-specific primers (GSPs). The starting template is cDNA generated from viral RNA using an oligo (dT) primer. The dT primer has two degenerate nucleotide positions at the 3’-end that position the primer at the start of the polyadenylated (polyA) tail. Based on the northern blot results, a cDNA library was generated using L RNA only. The GSPs (Table 2.2) were designed downstream from the putative start codon and upstream from the putative stop codon.

The 5’- and 3’-RACE PCR products were separated by agarose gel electrophoresis. Single bands of approximately 350 and 390 bp, were obtained respectively, for 5’-RACE for UL146 and UL139 (Figure 4.3).

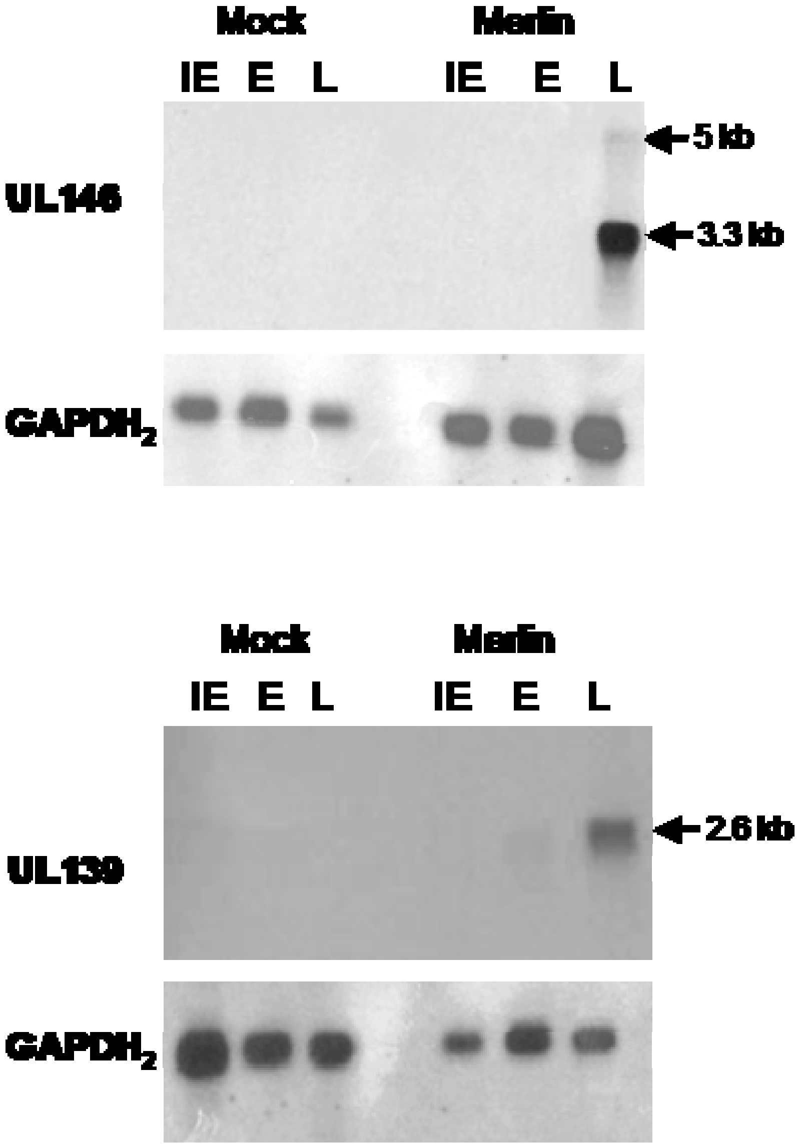


Figure 4.2 Northern blot analysis of UL146 and UL139

IE, E and L RNA was prepared from HFFFs infected with HCMV Merlin or mock infected. RNA was gel electrophoresed and transferred to a membrane, which was hybridised with a UL146-specific or a UL139-specific DIG-labelled ssRNA probe. Duplicate membranes were hybridised with a DIG-labelled, GAPDH$_2$-specific ssRNA probe.

Single bands were also obtained for 3'-RACE for UL146 and UL139, and were 3.1 kbp and 2.4 kbp in size, respectively (Figure 4.3). These PCR products were purified and cloned into pGem-T and a number of clones were sequenced for each. The sequences were compared with the Merlin sequence using Blast and the 5'- and 3'-ends were defined.

Table 4.1: 5'-ends of UL146 and UL139 mRNAs

ORF	*Position of 5'-end*[a]	*Putative TATA element (5'-3')*	*Position of TATA box*[a]	*Number of clones*[b]	*Position of ATG codon*[a]
UL146	181365/6	TACTTA	181395-181390	22	181292
UL139	187003	TATAAT	187035-187030	10	186878

[a] With reference to RefSeq accession NC_006273.2 (HCMV strain Merlin)
[b] Total number of clones corresponding to the mapped 5'-end

4.3.1 Sequences of 5'-ends

A single 5'-RACE band for UL146 suggests a single 5'-end and, indeed, sequencing of 22 clones for UL146 confirmed this (Table 4.1). However, as there is a G residue at this position there is some ambiguity as to whether the 5'-end is at position 181365 or 181366 or both. This indicates that transcription initiates 72-73 bp upstream of the UL146 initiation codon, and 27-28 bp downstream from a potential TATA element (TACTTA). Figure 4.5 shows the location of the 5'- end of the UL146 mRNA and the putative TATA element.

Similarly, a single 5'-RACE band for UL139 suggests a single 5'-end, and, indeed, sequencing of ten clones for UL139 confirmed this (Table 4.1). This indicates that transcription initiates 125 bp upstream of the UL139 start codon, and 27 bp downstream from a potential TATA element (TATAAT). Figure 4.6 shows the nucleotide location of the 5'-end of the UL139 mRNA and the putative TATA element.

4.3.2 Sequences of 3'-ends

3'-RACE of UL146 yielded a single band and 11 clones mapped the 3'-end to the same position, with the polyA signal downstream from the UL132 stop codon (Figure 4.5, Table 4.2).

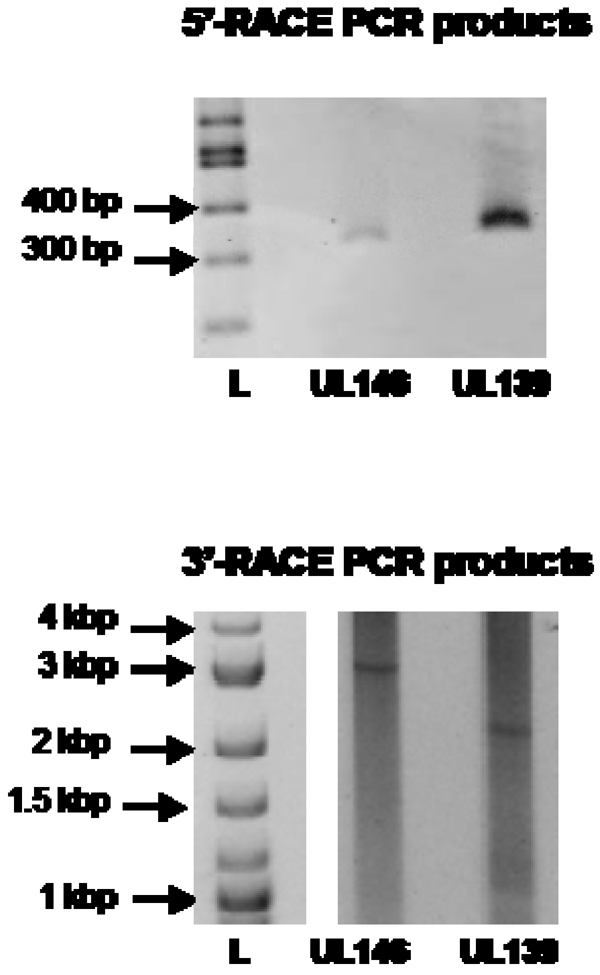


Figure 4.3 5'- and 3'-RACE of UL146 and UL139

5'- and 3'-cDNA was produced from L mRNA and 5'- and 3'-RACE was performed using GSPs (Table 2.2) and a universal primer. Representative 5'- and 3'-RACE PCR products containing UL139 and UL146 were electrophoresed on a 1% agarose gel and the gel was stained with ethidium bromide. L is a 2-log DNA ladder (New England Biolabs). Note that the 3'RACE lanes were taken from gel with additional samples in adjacent lanes.

This is consistent with the size of the 3'-RACE product obtained (Figure 4.3). The distance from the 5'-end to the 3'-end as mapped by RACE (3216 bp) is consistent with the size of the UL146 transcript (~3.3 kb) estimated from northern blot analysis (Figure 4.2). 3'-RACE of UL139 yielded a single band (Figure 4.3) and eight clones mapped the 3'-end to the same position, with the polyA signal downstream from the UL141 stop codon (Figure 4.6, Table 4.2). This is consistent with the size of the 3'-RACE product (Figure 4.3). The distance from the 5'-end to the 3'-end as mapped by RACE (2640 bp) correlates with the 2.6 kb transcript detected by northern blot analysis (Figure 4.2).

4.3.3 Mapping the 3'-ends of UL140 and UL141

Transcription of UL139 initiates upstream of UL139 and continues though UL139 and adjacent protein-coding region of UL140 and UL141. This suggests that UL140 and UL141 are 3'-coterminal with UL139. 3'-RACE was performed for UL140 and UL141 (using GSPs shown in Table 2.2) to investigate this possibility.

3'-RACE of UL140 yielded three bands, a major band 1.8 kbp in size and two minor bands, 0.65 and 0.5 kbp in length (Figure 4.4). Two clones derived from the 1.8 kbp band mapped the 3'-end of the UL140 transcript to the same position as that determined for UL139, confirming that UL140 is 3'-cotranscribed with UL139 (Figure 4.6).

3'-RACE of UL141 yielded three bands, a minor band 2.2 kbp in size, and a doublet band of approximately 0.8 kbp in size (Figure 4.4). Minor 3'-RACE bands obtained for both UL140 and UL141 mapped to polyA tracts located within RL5A and are likely a result of mispriming.

Table 4.2: 3'-ends of UL146, UL139, UL140 and UL141 mRNAs

ORF	*Position of stop codon*[a]	*Position of 3'-end*[a]	*PolyA signal*	*Position of polyA signal*[a]	*Distance*[b] *(bp)*	*Number of clones*[b]
UL146	180932	178149	AATAAA	178174-178169	3216	11
UL139	186464	184363	AATAAA	184382-184377	2640	8
UL140	185707	184363	AATAAA	184382-184377	ND	2
UL141	184398	184363	AATAAA	184382-184377	ND	5

[a] With reference to RefSeq accession NC_006273.2 (HCMV strain Merlin)

Distance[b] is from the mapped 5'-end to the mapped 3'-end (from RACE)

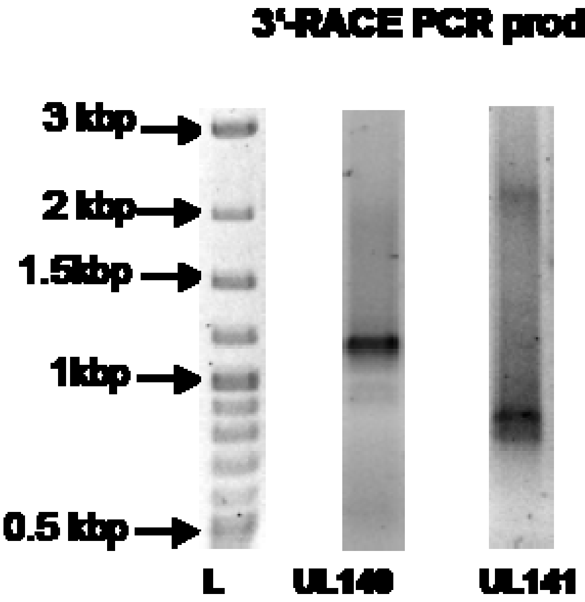


Figure 4.4 3'-RACE of UL140 and UL141

3'-cDNA was produced from L mRNA and 3'-RACE was performed using GSPs (Table 2.2) and a universal primer. Representative 3'-RACE PCR products containing UL140 and UL141 were electrophoresed on a 1% agarose gel and the gel was stained with ethidium bromide. L is a 2-log DNA ladder (New England Biolabs).Note lanes taken from gel with additional samples in adjacent lanes.

```
GGGGGGTACTTATCGGGAATTGATGTGTCATGGACGCAGTTTTGAGTGATTTTCCGGGAATACCGGATATTACGAATTATTGGTAGTGACGTAAATAATA  181301
        UL146
         M  R  L  I  F  G  A  L  I  I  S  L  T  Y  M  Y  Y  Y  E  V  H  G  T  E  L  R  C  K  C  L  D
AAATTATAATGCGATTAATTTTTGGTGCGTTGATTATTTCTTTAACGTATATGTATTATTATGAAGTGCATGGAACGGAATTACGCTGCAAATGTCTTGA  181201

  G  K  K  L  P  P  K  T  I  M  L  G  N  F  W  F  H  R  E  S  G  G  P  R  C  N  N  N  E  Y  F  L  Y
TGGTAAAAAACTGCCGCCCAAAACAATTATGTTGGGTAATTTTTGGTTTCATCGCGAATCTGGTGGTCCCAGATGCAATAACAATGAATATTTCTTGTAT  181101

 L  G  G  G  K  K  H  G  P  G  V  C  L  S  P  H  H  P  F  S  K  W  L  D  K  R  N  D  N  R  W  Y  N
CTAGGCGGAGGAAAAAAACATGGACCTGGAGTATGTTTATCGCCCCATCACCCTTTTTCAAAATGGCTAGACAAACGCAACGATAACAGGTGGTATAATG  181001

V  N  V  T  R  Q  P  E  R  G  P  G  K  I  T  V  T  L  V  G  L  K  E  -
TTAATGTAACAAGACAACCGGAACGAGGGCCGGGAAAAATAACTGTAACCCTAGTAGGTCTGAAGGAATAATATTTAGTATATATTTTAAACAGACAAGT  180901
                             UL147
                                M  L  L  T  W  L  H  H  P  I  L  N  S  R  I  K  L  L  S  V  R  Y  L
TTGTTAGAGCAGAAAATATCATGTTTTCAATATGTTGCTAACATGGTTACATCATCCGATTCTGAATTCGCGCATTAAACTTTTATCGGTACGATACCTG  180801

 S  L  T  A  Y  M  L  L  A  I  C  P  I  A  V  R  L  L  E  L  E  D  Y  D  K  R  C  R  C  N  N  Q  I
TCATTGACCGCATATATGTTACTTGCCATATGTCCCATAGCCGTCCGTCTTTTAGAACTAGAAGATTACGACAAGCGGTGTCGCTGTAATAACCAAATTC  180701

L  L  N  T  L  P  V  G  T  E  L  L  K  P  I  A  A  S  E  S  C  N  R  Q  E  V  L  A  I  L  K  D  K  G
TGTTGAATACCCTGCCGGTCGGAACCGAATTGCTTAAGCCAATCGCAGCGAGCGAAAGCTGCAATCGTCAGGAAGTGCTGGCTATTTTAAAGGACAAGGG  180601

  T  K  C  L  N  P  N  A  Q  A  V  R  R  H  I  N  R  L  F  F  R  L  V  L  D  E  E  Q  R  I  Y  D  V
CACCAAGTGTCTCAATCCTAACGCGCAAGCCGTGCGTCGTCACATCAACCGGCTATTTTTTCGGTTAGTCTTAGACGAGGAACAACGCATTTACGACGTA  180501

 V  S  T  N  I  E  F  G  A  W  P  V  P  T  A  Y  K  A  F  L  W  K  Y  A  K  K  L  N  Y  H  Y  F  R
GTGTCTACAAATATTGAGTTCGGTGCCTGGCCAGTCCCTACGGCCTACAAAGCCTTTCTCTGGAAATACGCCAAGAAACTGAATTACCACTACTTTAGAC  180401
             UL147A
L  R  W  -    M  S  L  F  Y  R  A  V  A  L  G  T  L  S  A  L  V  W  Y  S  T  S  I  L  A  E  I  N  E
TGCGCTGGTGATCATGTCCCTATTTTACCGTGCGGTAGCTCTGGGCACACTAAGCGCTCTGGTGTGGTACAGCACTAGTATCCTCGCAGAGATTAACGAA 180301

 N  S  C  S  S  S  S  V  D  H  E  D  C  E  E  P  D  E  I  V  R  E  E  Q  D  Y  R  A  L  L  A  F  S
AATTCCTGCTCCTCATCTTCTGTGGACCACGAAGATTGCGAGGAACCGGACGAGATCGTTCGCGAAGAGCAAGACTATCGGGCTCTGCTGGCCTTTTCCC 180201

L  V  I  C  G  T  L  L  V  T  C  V  I  -
TAGTGATTTGCGGTACGCTCCTCGTCACTTGTGTGATCTGAGACGTCATGCTGGTAGCGTTTATGAGTCGGGCGGTGGCCGGCACGCCGCATTTCCTAAC 180101
        UL148
          M  L  R  L  L  F  T  L  V  L  L  A  L  Y  G  P  S  V  D  A  S  R  D  Y  V  H  V  R  L  L
CCGCGCAGCATGTTGCGCTTGCTGTTCACGCTCGTACTGCTGGCCCTCTACGGACCGTCTGTCGACGCTAGCCGCGACTATGTGCATGTTCGACTACTGA 180001

S  Y  R  G  D  P  L  V  F  K  H  T  F  S  G  V  R  R  P  F  T  E  L  G  W  A  A  C  R  D  W  D  S  M
GCTACCGAGGCGACCCCCTGGTCTTCAAGCACACTTTCTCGGGTGTGCGTCGACCCTTCACCGAGCTAGGCTGGGCTGCGTGTCGCGACTGGGACAGTAT 179901

  H  C  T  P  F  W  S  T  D  L  E  Q  M  T  D  S  V  R  R  Y  S  T  V  S  P  G  K  E  V  T  L  Q  L
GCATTGCACGCCCTTCTGGTCTACCGATCTGGAGCAGATGACCGACTCGGTGCGGCGTTACAGCACGGTGAGCCCCGGCAAGGAAGTGACGCTTCAGCTT 179801

 H  G  N  Q  T  V  Q  P  S  F  L  S  F  T  C  R  L  Q  L  E  P  V  V  E  N  V  G  L  Y  V  A  Y  V
CACGGGAACCAAACCGTACAGCCGTCGTTTCTAAGCTTTACGTGCCGCCTGCAGCTAGAACCCGTGGTGGAAAATGTTGGCCTCTACGTGGCCTACGTGG 179701

V  N  D  G  E  R  P  Q  Q  F  F  T  P  Q  V  D  V  V  R  F  A  L  Y  L  E  T  L  S  R  I  V  E  P  L
TCAACGACGGTGAACGCCCACAACAGTTTTTTACACCGCAGGTAGACGTGGTACGCTTTGCTCTATATCTAGAAACGCTCTCCCGGATCGTGGAACCGTT 179601

  E  S  G  R  L  T  V  E  F  D  T  P  D  L  A  L  A  P  D  L  V  S  S  L  F  V  A  G  H  G  E  T  D
AGAATCAGGTCGCCTGACAGTGGAATTTGATACGCCTGACCTAGCTCTGGCGCCCGATTTAGTAAGCAGCCTCTTCGTGGCCGGACACGGCGAGACCGAC 179501

 F  Y  M  N  W  T  L  R  R  S  Q  T  H  Y  L  E  E  M  A  L  Q  V  E  I  L  K  P  R  G  V  R  H  R
TTTTACATGAACTGGACGCTGCGTCGCAGTCAGACCCACTACCTGGAGGAGATGGCCTTACAGGTGGAGATTCTAAAGCCCCGCGGCGTACGTCACCGCG 179401

A  I  I  H  H  P  K  L  Q  P  G  V  G  L  W  I  D  F  C  V  Y  R  Y  N  A  R  L  T  R  G  Y  V  R  Y
CTATTATCCACCATCCGAAGCTACAACCGGGCGTTGGCTTGTGGATAGATTTCTGCGTGTACCGCTACAACGCGCGCCTGACCCGTGGCTACGTACGATA 179301

  T  L  S  P  K  A  R  L  P  A  K  A  E  G  W  L  V  S  L  D  R  F  I  V  Q  Y  L  N  T  L  L  I  T
CACCCTGTCACCGAAAGCGCGCTTGCCCGCAAAAGCAGAGGGTTGGCTGGTGTCACTAGACAGATTCATCGTGCAGTACCTCAACACATTGCTGATTACA 179201

 M  M  A  A  I  W  A  R  V  L  I  T  Y  L  V  S  R  R  R  -
ATGATGGCGGCGATATGGGCTCGCGTTTTGATAACCTACCTGGTGTCGCGGCGTCGGTAGAGGCTTGCGGAAACCACGTCCTCGTCACACGTCGTTCGCG 179101
                                   UL132
                                     M  P  A  P  R  G  P  L  R  A  T  F  L  A  L  V  A  F  G  L  L
GACATAGCAAGAAATCCACGTCGCCACGTCTCGAGAATGCCGGCCCCGCGGGGTCCCCTTCGCGCAACATTCCTGGCCCTGGTCGCGTTCGGGTTGCTGC 179001

L  Q  I  D  L  S  D  V  T  N  V  T  S  S  T  K  V  P  T  S  T  S  N  R  N  S  V  D  N  A  T  S  S  G
TTCAGATAGACCTCAGCGACGTTACGAATGTGACCAGCAGCACAAAAGTCCCTACTAGCACCAGCAACAGAAATAGCGTCGACAACGCCACGAGTAGCGG 178901

  P  T  T  G  I  N  M  T  T  T  H  E  S  S  V  H  N  V  R  N  N  E  I  M  K  V  L  A  I  L  F  Y  I
ACCCACGACCGGGATCAACATGACCACCACCCACGAGTCTTCCGTTCACAACGTGCGCAATAACGAGATCATGAAAGTGCTGGCTATCCTCTTCTACATC 178801

 V  T  G  T  S  I  F  S  F  I  A  V  L  V  A  V  V  Y  S  S  C  C  K  H  P  G  R  F  R  F  A  D  E
GTGACAGGCACCTCCATTTTCAGCTTCATAGCGGTACTGGTCGCGGTAGTTTACTCCTCGTGTTGCAAGCACCCGGGTCGCTTTCGTTTCGCCGACGAAG 178701

E  A  V  N  L  L  D  D  T  D  D  S  G  G  S  S  P  F  G  S  G  S  R  R  G  S  Q  I  P  A  G  F  C  S
AAGCCGTCAACCTGTTGGACGACACGGACGACAGTGGCGGCAGCAGCCCGTTTGGCAGCGGTTCCCGACGAGGTTCTCAGATCCCCGCCGGATTTTGTTC 178601

  S  S  P  Y  Q  R  L  E  T  R  D  W  D  E  E  E  E  A  S  A  A  R  E  R  M  K  H  D  P  E  N  V  I
CTCGAGCCCTTATCAGCGGTTGGAAACTCGGGACTGGGACGAGGAGGAGGAGGCGTCCGCGGCCCGCGAGCGCATGAAACATGATCCTGAGAACGTCATC 178501

 Y  F  R  K  D  G  N  L  D  T  S  F  V  N  P  N  Y  G  R  G  S  P  L  T  I  E  S  H  L  S  D  N  E
TATTTCAGAAAGGATGGCAACTTGGACACGTCGTTCGTGAATCCCAATTATGGGAGAGGCTCGCCTTTGACCATCGAATCTCACCTCTCGGACAATGAGG 178401

E  D  P  I  R  Y  Y  V  S  V  Y  D  E  L  T  A  S  E  M  E  E  P  S  N  S  T  S  W  Q  I  P  K  L  M
AGGACCCCATCAGGTACTACGTTTCGGTGTACGATGAACTGACCGCCTCGGAAATGGAAGAACCTTCGAACAGCACCAGCTGGCAGATTCCCAAACTAAT 178301

  K  V  A  M  Q  P  V  S  L  R  D  P  E  Y  D  -
GAAAGTTGCCATGCAACCCGTCTCGCTCAGAGATCCCGAGTACGACTAGGCTTTTTTTTTTGTCTTTCGGTTCCAACTCTTTCCCCGCCCCATCACCTCG 178201

CCTATACTATGTGTATGATGTCTCATAATAAAGCTTTCTTTCTCAGTCTGCAACATGCGG   178140
```

Figure 4.5 Location of the 5’- and 3’-ends of the UL146 mRNA

Figure 4.5 Location of the 5’- and 3’-ends of the UL146 mRNA
The nucleotide sequence of the region of the genome containing UL146, UL147, UL147A, UL148 and UL132 and the encoded amino acids are shown, with the Merlin coordinates on the right. All genes are shown in reverse orientation with respect to the Merlin genome, therefore they are oriented left to right.

The position of the 5’-end of UL146 is highlighted in pink (181365/6). The 5’-RACE primer is underlined and orientated right to left. The putative TATA element (TACTTA) is highlighted in red and the initiation codon is highlighted in blue. The UL146 stop codon is highlighted in green. The initiation and stop codons of UL147, UL147A, UL148 and UL132 are also highlighted blue and green, respectively.

The position of the 3’-end of UL146 is highlighted in plum. The 3’-RACE primer is highlighted in yellow and orientated left to right. The putative polyA signal (AATAAA) is highlighted in red.

```
GTAGTGGAAATTTTTACGTCATTGGGAAACCCCAGAATGAAAGAGTATAATGTGCACATCACCGGGGGTTCCCTTTCAGTACGAATGTACACAACGCGGG 186981

TTACATTACGATAAACTTTCCGGTAAAACGATGCCGATACAGCGTATATAACGCTGATTGTCACGACAAAGGGGTTCGTATATCAATTATATAGTAACGA 186881
  UL139
   M  L  W  I  L  V  L  F  A  L  A  A  S  A  S  E  T  T  T  G  T  S  S  N  S  S  Q  S  T  S  A  G  T
ACATGCTGTGGATATTAGTTTTATTTGCACTTGCCGCATCGGCGAGTGAAACCACTACAGGTACCAGCTCTAATTCCAGTCAATCTACTTCTGCTGGTAC 186781

  T  N  T  T  T  P  S  T  A  C  I  N  A  S  N  G  S  D  L  G  A  P  Q  L  A  L  L  A  A  S  G  W  T
CACTAACACGACTACACCATCGACAGCATGTATTAATGCTTCTAACGGCAGTGATTTGGGGGCGCCACAGCTCGCGCTACTTGCCGCTAGCGGCTGGACA 186681

 L  S  G  L  L  L  I  F  T  C  C  L  C  C  F  W  L  V  R  K  V  C  S  C  C  G  N  S  S  E  S  E  S
TTATCTGGACTCCTTCTCATATTTACTTGCTGCCTTTGCTGTTTTTGGCTAGTACGTAAAGTCTGCAGCTGCTGCGGCAACTCCTCCGAGTCAGAGAGCA 186581

K  A  T  H  A  Y  T  N  A  A  F  T  S  S  D  A  T  L  P  M  G  T  T  G  S  Y  T  P  P  Q  D  G  S  F
AAGCCACTCACGCGTACACCAATGCCGCATTCACTTCTTCCGATGCAACGTTACCCATGGGCACTACAGGGTCGTACACTCCCCCACAGGACGGCTCATT 186481

  P  P  P  P  R  -
TCCACCTCCGCCTCGGTGACGCAGGCTAAACCGAAACCAACGTTGAACTTGACGCGGTTTCGGAAAGCCTGAGACGTCACTTTCACAATGACGTTCGTAG 186381

ACACGTTGATCATAAAACACCGTAGAGGCTAAGGCTTCGGTAGGGAGACACCTCAACTGTTCCTGATGAGCACCCGCGCTCTCATCTCTTCAGACTTGTC 186281
UL140
 M  T  P  A  Q  T  N  G  T  T  T  V  H  P  H  G  A  K  N  G  S  G  G  S  A  L  P  T  L  V  V  F  G
ATGACCCCCGCTCAGACTAACGGCACTACCACCGTGCACCCGCACGGCGCAAAAAACGGCAGCGGCGGTAGTGCCCTGCCGACCCTCGTCGTTTTCGGCT 186181

F  I  V  T  L  L  F  F  L  F  M  L  Y  F  W  N  N  D  V  F  R  K  L  L  R  C  A  W  I  Q  R  C  C  D
TCATCGTTACGCTACTTTTCTTTCTCTTTATGCTCTACTTTTGGAACAACGACGTGTTCCGTAAGCTGCTTCGCTGCGCTTGGATCCAGCGCTGCTGCGA 186081

  R  F  D  A  W  Q  D  E  V  I  Y  R  R  P  S  R  R  S  Q  S  D  D  E  S  R  T  N  S  V  S  S  Y  V
CCGCTTCGACGCGTGGCAAGACGAGGTCATCTACCGTCGTCCATCACGTCGTTCCCAAAGCGACGACGAGAGTCGTACTAACAGCGTGTCATCGTACGTT 185981

 L  L  S  P  A  S  D  G  G  F  D  N  P  A  L  T  E  A  V  D  S  V  D  D  W  A  T  T  S  V  F  Y  A
CTTTTATCACCCGCGTCCGATGGCGGTTTTGACAACCCGGCACTGACAGAAGCCGTCGACAGCGTGGACGACTGGGCGACCACCTCGGTTTTTTACGCCA 185881

T  S  D  E  T  A  D  T  E  R  R  D  S  Q  Q  L  L  I  E  L  P  P  E  P  L  P  P  D  V  V  A  A  M  Q
CGTCCGACGAAACGGCGGACACCGAACGCCGAGATTCGCAGCAACTGCTCATCGAGCTTCCGCCGGAGCCGCTCCCACCCGATGTGGTAGCGGCCATGCA 185781

  K  A  V  K  R  A  V  Q  N  A  L  R  H  S  H  D  S  W  Q  L  H  Q  T  L  -
GAAAGCGGTGAAACGCGCTGTACAAAACGCGCTACGCCACAGCCACGACTCTTGGCAGCTTCATCAGACCCTGTGACGCAGATAAACGTTCCTTCTTAAA 185681

CATCCGAGGTAGCAATGAGACAGGTCGCGTACCGCCGGCGACGCGAGAGTTCCTGCGCGGTGCTGGTCCACCACGTCGGCCGCGACGGCGAGGGAGAGGC 185581

AGCAAAAAAGACCTGTAAAAAAACCGGACGCTCAGTTGCGGGCATCCCGGGCGAGAAGCTGCGTCGCACGGTGGTCACCACCACGCCGGCCCGACGTTTG 185481
                                                                    UL141
                                                                     M  C  R  R  E  S  L  R  T  L  P
AGCGGCCGACACACGGAGCAGGAACAGGCGGGCAGCGTCTCTGCGAAAAAGGGAAGAAAAGAATCATCATGTGCCGCCGGGAGTCGCTCCGAACTCTGCC 185381

  W  L  F  W  V  L  L  S  C  P  R  L  L  E  Y  S  S  S  S  F  P  F  A  T  A  D  I  A  E  K  M  W  A
GTGGCTGTTCTGGGTGCTGTTGAGCTGCCCGCGACTCCTCGAATATTCTTCCTCTTCGTTCCCCTTCGCCACCGCTGACATCGCCGAAAAGATGTGGGCC 185281

 E  N  Y  E  T  T  S  P  A  P  V  L  V  A  E  G  E  Q  V  T  I  P  C  T  V  M  T  H  S  W  P  M  V
GAGAACTATGAGACCACGTCGCCGGCGCCGGTGTTGGTCGCCGAGGGAGAGCAAGTTACCATCCCCTGCACGGTCATGACACACTCCTGGCCCATGGTTT 185181

S  I  R  A  R  F  C  R  S  H  D  G  S  D  E  L  I  L  D  A  V  K  G  H  R  L  M  N  G  L  Q  Y  R  L
CCATTCGCGCACGTTTCTGTCGTTCCCACGACGGCAGCGACGAGCTCATCCTGGACGCCGTCAAAGGCCATAGGCTGATGAATGGACTTCAATACCGCCT 185081

  P  Y  A  T  W  N  F  S  Q  L  H  L  G  Q  I  F  S  L  T  F  N  V  S  T  D  T  A  G  M  Y  E  C  V
GCCGTACGCCACTTGGAATTTCTCGCAGTTGCATCTCGGCCAAATATTCTCGCTGACTTTCAACGTATCGACGGACACGGCCGGCATGTACGAATGCGTG 184981

 L  R  N  Y  S  H  G  L  I  M  Q  R  F  V  I  L  T  Q  L  E  T  L  S  R  P  D  E  P  C  C  T  P  A
CTGCGCAACTATAGCCACGGCCTCATCATGCAACGCTTCGTAATTCTGACGCAACTGGAGACGCTCAGCCGGCCCGACGAACCTTGCTGCACGCCGGCGT 184881

L  G  R  Y  S  L  G  D  Q  I  W  S  P  T  P  W  R  L  R  N  H  D  C  G  M  Y  R  G  F  Q  R  N  Y  F
TAGGTCGCTACTCGCTGGGAGACCAGATCTGGTCGCCGACGCCCTGGCGTCTACGGAATCACGACTGCGGGATGTACCGCGGTTTTCAACGCAACTACTT 184781

  Y  I  G  R  A  D  A  E  D  C  W  K  P  A  C  P  D  E  E  P  D  R  C  W  T  V  I  Q  R  Y  R  L  P
CTATATCGGCCGCGCCGACGCCGAGGATTGCTGGAAACCCGCATGTCCGGACGAGGAACCCGACCGCTGTTGGACAGTGATACAGCGTTACCGGCTCCCC 184681

 G  D  C  Y  R  S  Q  P  H  P  P  K  F  L  P  V  T  P  A  P  P  A  D  I  D  T  G  M  S  P  W  A  T
GGCGACTGCTACCGTTCGCAGCCACACCCGCCGAAATTTTTACCGGTGACGCCAGCACCGCCGGCCGACATAGACACCGGGATGTCTCCCTGGGCCACTC 184581

R  G  I  A  A  F  L  G  F  W  S  I  F  T  V  C  F  L  C  Y  L  C  Y  L  Q  C  C  G  R  W  C  P  T  P
GGGGAATCGCGGCATTTTTGGGATTTTGGAGTATTTTCACCGTATGTTTCCTATGCTACCTGTGTTACCTGCAGTGCTGTGGACGCTGGTGCCCCACGCC 184481

  G  R  G  R  R  G  G  E  G  Y  R  R  L  P  T  Y  D  S  Y  P  G  V  K  K  M  K  R  -
GGGAAGGGGACGACGAGGCGGTGAGGGCTATCGACGCCTACCGACTTACGATAGTTACCCCGGTGTTAAAAAGATGAAGAGGTGAGAACACGCATAAAAT 184381


AAAAAAATAAGATGTTAAAAAAATGCAGTGTGTGAAATGTGAATAGTGTGATTAAAATATGCGGATTGAAT 184311
```

Figure 4.6 Location of the 5’- and 3’-ends of the UL139 mRNA
(continued overleaf)

Figure 4.6 Location of the 5'- and 3'-ends of the UL139 mRNA

The nucleotide sequence of the region of the genome containing UL139, UL140 and UL141 and the encoded amino acids are shown, with the Merlin coordinates on the right. All genes are shown in reverse orientation with respect to the Merlin genome, therefore they are oriented left to right.

The position of the 5'-end of UL139 is highlighted in pink (187003). The 5'-RACE primer is underlined and orientated right to left. The putative TATA element (TATAAT) is highlighted in red and the initiation codon is highlighted in blue. The UL139 stop codon is highlighted in green. The initiation and stop codons of UL140 and UL141 are also highlighted blue and green, respectively.

The position of the 3'-ends of UL139, UL140 and UL141 mRNA are highlighted in plum. The 3'-RACE primers are highlighted in yellow and orientated left to right. The putative polyA signal (AATAAA) is highlighted in red.

Five clones mapped the 3'-end to the same position as that determined for UL139, with the polyA signal downstream of the UL141 stop codon, confirming that UL141 uses the same polyA signal as UL139 and UL140 (Figure 4.6).

4.4 Construction of recombinant adenoviruses expressing tagged UL139 variants

As shown in Chapter 1, seven of the eight UL139 variants share a small region of similarity with CD24. Similarity with CD24 could indicate a role in immune modulation for UL139, which is an intriguing prospect. Based on potential roles for UL139 in virus pathogenesis, immune modulation and tissue tropism, an initial characterisation of the UL139 protein was undertaken, with a focus on the apparent masses of the proteins produced by different genotypes. As described in Chapter 1 (Section 1.10), the primary translation product of UL139 from N to C terminus consists of a signal peptide sequence, the CD24-related region, a hypervariable region, a transmembrane anchor, and a highly conserved cytoplasmic tail (Figure 4.7). Thus, the ectodomain of the mature protein consists of the CD24-related domain (but not in G5) followed by the hypervariable region. As emphasized in Figure 4.7, the ectodomain is rich in S and T residues, which are potentially subject to *O*-glycosylation, and NXS/NXT motifs, which are potentially subject to *N*-glycosylation. The hypothesis was that the mature UL139 protein is highly glycosylated and therefore has an apparent mass much greater than that predicted from the aa sequence. Moreover, apparent mass would vary among genotypes.

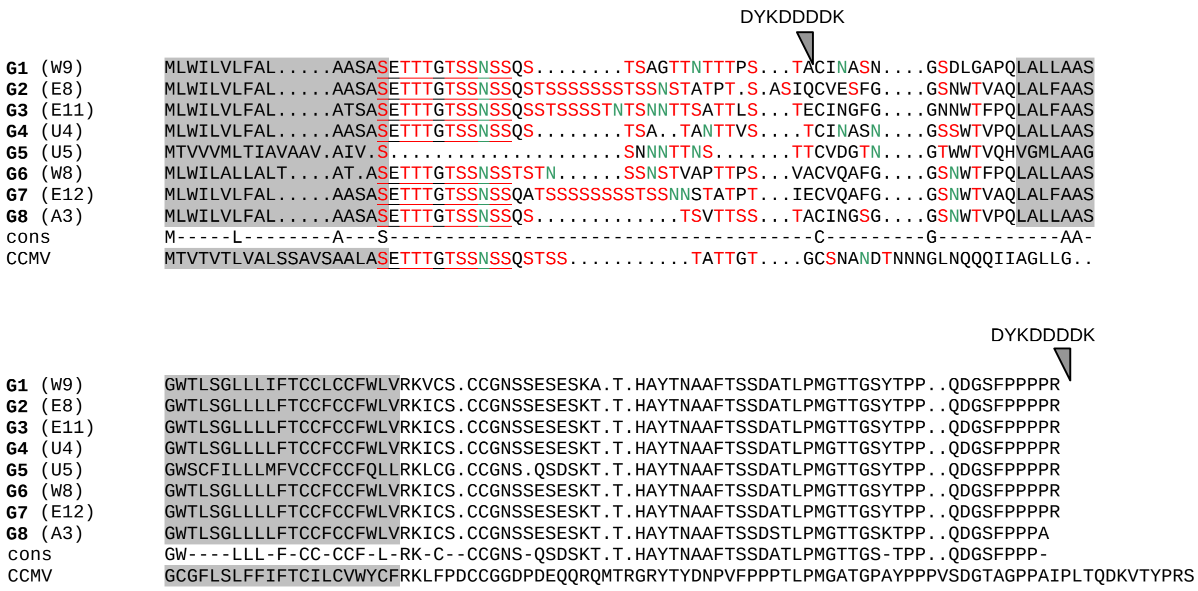
The UL139 protein was over-expressed in a RAD vector under the control of the HCMV MIE promoter. As no UL139 antibody was available, the proteins were FLAG-tagged to facilitate detection by immunofluorescence and immunoblot. The FLAG-tag was selected due its small size (~1 kDa) and the availability of a commercial highly specific antibody. Constructs were designed with the FLAG-tag at the C terminus or internally in three different genotypes: G1, G2 and G5 (Figure 4.7). G1 is found in Merlin, which is being used increasingly as a laboratory strain. G2 differs considerably from G1 and yet still contains the region of similarity to CD24. G5 is the only UL139 genotype that does not contain the region of similarity to CD24.

The insertion sites for the FLAG-tag (Figure 4.7) were either internal, downstream of the hypervariable region, or at the C-terminus. The three internally tagged variants derived for different genotypes are referred to as RADUL139G1int, RADUL139G2int and RADUL139G5int. The three C terminally tagged variants are referred to as RADUL139G1ter, RADUL139G2ter and RADUL139G5ter.

The six UL139 FLAG-tagged sequences were provided commercially by genesynthesis (Genscript Corporation) as inserts in the multiple cloning site of the plasmid pUC57. Briefly, genesynthesis is the chemical synthesis of oligonucleotides in a manner similar to the synthesis of primers (using phosphoramidites, normal nucleotides which have protection groups, allowing specific addition of each nucleotide) followed by ligation (Gupta *et al.*, 1968). They were designed to contain regions or ‘arms’ of sequence homology with the adenovirus vector (pAL942) at the 5’- and 3’-ends. Each tagged UL139 variant was purified and electroporated into *E. coli* SW102 cells (Chapter 2, Section 2.25). SW102 cells contain pAL942, the adenovirus BAC vector with which the tagged UL139 variants were to recombine. The cells were recovered and plated onto ampicillin plates. A total of 40-50 colonies were tested directly by PCR, which was performed using primers (PMV100f and PMV100r, Table 2.2) selected within the ‘arms’ of homology. Colonies containing the UL139 inserts were inoculated into L-broth containing appropriate antibiotics. The BAC DNAs were purified and sequenced. Five of the UL139 variants (RADUL139G1int, RADUL139G1ter, RADUL139G2int, RADUL139G2ter and RADUL139G5ter) successfully recombined with pAL942 as verified by sequencing. BAC DNA was excised to produce the adenovirus genome containing the relevant insert, and viral DNA was purified. The purified viral DNA was then transfected into HEK 293 cells for large-scale RAD production for use in further experiments.

4.5 Detection of UL139 FLAG-tagged variants by immunoblot

To confirm that the FLAG-tagged UL139 variants were expressed and to examine the apparent masses of the proteins, HFFF-2 cells were infected with each RAD at an m.o.i. of 100 p.f.u./ml and harvested 72 hours p.i. An empty RAD (RAD942) was used as a negative control. Based on the otherwise unprocessed aa



```
G1 (W9)     MLWILVLFAL.....AASASETTTGTSSNSSQS........TSAGTTNTTTPS...TACINASN....GSDLGAPQLALLAAS
G2 (E8)     MLWILVLFAL.....AASASETTTGTSSNSSQSTSSSSSSSTSSNSTATPT.S.ASIQCVESFG....GSNWTVAQLALFAAS
G3 (E11)    MLWILVLFAL.....ATSASETTTGTSSNSSQSSTSSSSTNTSNNTTSATTLS...TECINGFG....GNNWTFPQLALFAAS
G4 (U4)     MLWILVLFAL.....AASASETTTGTSSNSSQS........TSA..TANTTVS....TCINASN....GSSWTVPQLALLAAS
G5 (U5)     MTVVVMLTIAVAAV.AIV.S.....................SNNNTTNS.......TTCVDGTN....GTWWTVQHVGMLAAG
G6 (W8)     MLWILALLALT....AT.ASETTTGTSSNSSTSTN......SSNSTVAPTTPS...VACVQAFG....GSNWTFPQLALLAAS
G7 (E12)    MLWILVLFAL.....AASASETTTGTSSNSSQATSSSSSSSSTSSNNSTATPT...IECVQAFG....GSNWTVAQLALFAAS
G8 (A3)     MLWILVLFAL.....AASASETTTGTSSNSSQS.............TSVTTSS...TACINGSG....GSNWTVPQLALLAAS
cons        M-----L--------A---S------------------------------------C---------G-----------AA-
CCMV        MTVTVTLVALSSAVSAALASETTTGTSSNSSQSTSS...........TATTGT....GCSNANDTNNNGLNQQQIIAGLLG..

G1 (W9)     GWTLSGLLLIFTCCLCCFWLVRKVCS.CCGNSSESESKA.T.HAYTNAAFTSSDATLPMGTTGSYTPP..QDGSFPPPPR
G2 (E8)     GWTLSGLLLLFTCCFCCFWLVRKICS.CCGNSSESESKT.T.HAYTNAAFTSSDATLPMGTTGSYTPP..QDGSFPPPPR
G3 (E11)    GWTLSGLLLLFTCCFCCFWLVRKICS.CCGNSSESESKT.T.HAYTNAAFTSSDATLPMGTTGSYTPP..QDGSFPPPPR
G4 (U4)     GWTLSGLLLLFTCCFCCFWLVRKICS.CCGNSSESESKT.T.HAYTNAAFTSSDATLPMGTTGSYTPP..QDGSFPPPPR
G5 (U5)     GWSCFILLLMFVCCFCCFQLLRKLCG.CCGNS.QSDSKT.T.HAYTNAAFTSSDATLPMGTTGSYTPP..QDGSFPPPPR
G6 (W8)     GWTLSGLLLLFTCCFCCFWLVRKICS.CCGNSSESESKT.T.HAYTNAAFTSSDATLPMGTTGSYTPP..QDGSFPPPPR
G7 (E12)    GWTLSGLLLLFTCCFCCFWLVRKICS.CCGNSSESESKT.T.HAYTNAAFTSSDATLPMGTTGSYTPP..QDGSFPPPPR
G8 (A3)     GWTLSGLLLLFTCCFCCFWLVRKICS.CCGNSSESESKT.T.HAYTNAAFTSSDSTLPMGTTGSKTPP..QDGSFPPPA
cons        GW----LLL-F-CC-CCF-L-RK-C--CCGNS-QSDSKT.T.HAYTNAAFTSSDATLPMGTTGS-TPP..QDGSFPPP-
CCMV        GCGFLSLFFIFTCILCVWYCFRKLFPDCCGGDPDEQQRQMTRGRYTYDNPVFPPPTLPMGATGPAYPPPVSDGTAGPPAIPLTQDKVTYPRS
```

Figure 4.7 FLAG-tag insertion sites in the eight UL139 genotypes

This is a version of Figure 3.8, which displays an amino acid sequence alignment of UL139 genotypes, G1-G8, with the S and T residues in red font, the N of NXS/T motifs in green font and the region of homology with CD24 underlined. The two alternative FLAG-tag insertion sites in UL139 G1, G2 and G5 are indicated by a grey triangle above the sequences with the amino acid sequence of the FLAG-tag shown (8 aa in length).

sequences minus the signal peptides using the Compute pI/Mw tool on www.expasy.org, RADUL139G1int and RADUL139G1ter were predicted to encode UL139 proteins with a mass of 13.5 kDa, RADUL139G2int and RADUL139G2ter of 14.7 kDa, and RADUL139G5ter of 12.3 kDa.

The results from immunoblotting are shown in Figure 4.8. RADUL139G1ter (lane 1) and RADUL139G1int (lane 2) are both 97 kDa in size, which is more than seven times their predicted size. RADUL139G2ter (lane 3), RADUL139G2int (lane 4) and RADUL139G5ter (lane 5) are ~110 kDa in size. Therefore all FLAG-tagged UL139 protein variants appear to be much larger than predicted. The amount of protein detected was much greater for RADUL139G2int, and therefore a lower exposure image was included in lane 4.

4.6 Discussion and future work

The first aim of this chapter was to determine the transcription kinetics and map the 5'- and 3'-ends of the UL146 and UL139 transcripts. As described in Section 1.4, there are three broad classes of HCMV gene expression; IE, E and L. The kinetic class of a particular HCMV gene can be determined by infecting cells in the presence of chemical inhibitors such as cycloheximide, a protein synthesis inhibitor that allows expression of IE genes, or phosphonoacetic acid (PAA), an inhibitor of viral DNA replication that allows expression of IE and E genes. All three classes of gene are expressed in the absence of these inhibitors. HFFF-2 cells were infected with HCMV Merlin, and IE, E and L genes were expressed using cycloheximide and PAA. Total RNA was extracted and electrophoresed, and the UL146 and UL139 transcripts were detected by northern blot. UL146 was transcribed as a 3.3 kb RNA with L kinetics whereas UL139 was expressed as a 2.6 kb RNA with E-L kinetics (Figure 4.2).

Having established the kinetic classes of UL146 and UL139, the 5'- and 3'-ends of the transcripts were mapped in L RNA using RACE. HCMV E and L genes usually have a promoter structure that contains a TATA box, which is a consensus sequence (based on TATAAA), located 25-35 nt upstream of the transcription start site (TSS). The 5'-end of the UL146 RNA is located at position 181365/181366 and the 3'-end is located at position 178149 on the Merlin genome (Figure 4.5). The location of the putative TATA element, 5'-end and 3'-

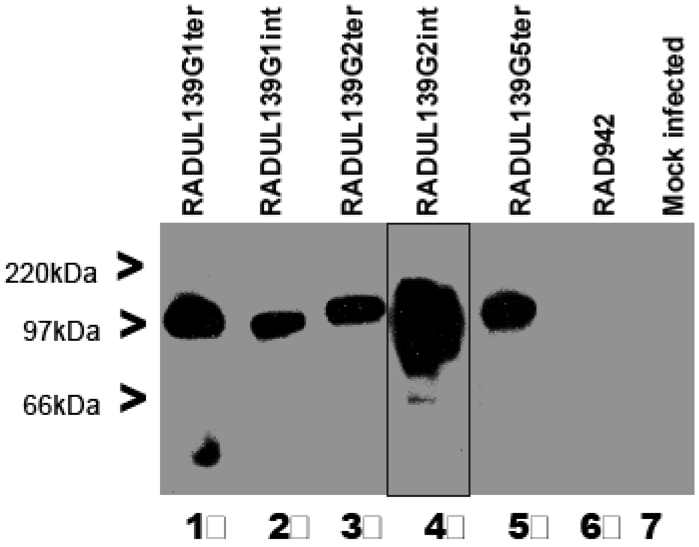


Figure 4.8 Immunoblot of FLAG-tagged UL139 variants

Protein extracts were prepared from HFFF-2 cells infected with RADs containing FLAG-tagged UL139 variants and electrophoresed on a SDS-PAGE gel. Loaded protein was detected using anti-FLAG primary antibody, followed by anti-mouse Ig HRP secondary antibody. Lane 4 displays an image taken at a lower exposure than the other lanes because it contained a much higher amount of protein.

end of UL146 are conserved in other HCMV strains. Transcription begins upstream of the UL146 start codon and continues through UL147, UL147A, UL148 and UL132 to a polyA signal downstream from the UL132 stop codon (Figure 4.5). It is likely that UL147, UL147A, UL148 and UL132 are 3'-coterminally transcribed with UL146, as the HCMV genome contains relatively few putative polyA signals therefore many genes share polyA signals (Wing and Huang, 1995). The putative transcriptional organisation for this locus is shown in Figure 4.1. During the course of this work Lurain *et al.* (2006) published data regarding transcription of UL146 and concluded that UL146 is indeed 3'-coterminally transcribed with adjacent genes. Further RACE experiments are required to map the 5'- and 3'-ends of UL147, UL147A, UL148 and UL132.

Penfold *et al.* (1999) found that UL146 was expressed with L kinetics, which is in agreement with the present study. However, Lurain *et al.* (2006) and Chambers *et al.* (1999) found that UL146 was expressed with E-L kinetics. Microarrays are unable to distinguish between overlapping transcripts, therefore discrepancies between the results of Chambers *et al.* (1999) and the present study may be a consequence of this technical limitation. The UL146 transcript detected by Lurain *et al.* (2006) with E-L kinetics (48 h p.i.) was faint, suggesting a low level of transcription. In contrast, the UL146 transcript detected 72 h p.i. was strong, which suggests a high level of transcription. Therefore, the discrepancy between their results and those of the present study may be a quantitative difference.

In contrast to UL146, UL139 is expressed with E-L kinetics (Figure 4.2), although levels of E transcription were very low. Transcription initiates upstream of the UL139 start codon and continues through UL140 and UL141 to a polyA signal downstream from the UL141 stop codon (Figure 4.6). From these data it was postulated that UL139 is 3'-coterminally expressed with UL140 and UL141, and, indeed, all three genes use the same polyA signal located downstream from the UL141 stop codon (Figure 4.6, Table 4.2). The putative transcriptional organisation for this locus based on the data obtained is displayed in Figure 4.1. The location of the putative TATA element, 5'-end and 3'-end of UL139 and 3'-ends of UL140 and UL141 are conserved in other HCMV strains. Moreover the putative TATA element for HCMV UL139 is also conserved in CCMV UL139, suggesting that transcription initiates within the gene known as CCMV UL139 to produce a protein beginning with the methionine aligned with the start codon of HCMV UL139 (Figure 3.13).

The second aim of this chapter was the initial characterisation of the UL139 protein, with a focus on the apparent mass generated from the different genotypes. Three UL139 genotypes (G1, G2 and G5) were chosen and FLAG-tagged internally or C-terminally. These tagged sequences were recombined into an adenovirus vector and five RADs (RADUL139G1int, RADUL139G1ter, RADUL139G2int, RADUL139G2ter and RADUL139G5ter) were generated.

Immunoblotting revealed that all five FLAG-tagged UL139 variants were expressed in infected HFFF-2 cells and all had molecular masses of ~97 kDa or more (Figure 4.8), much greater than their predicted masses of 12.3-14.7 kDa. It is possible that *N*-linked glycosyl groups (and other post-translational modifications) may account for the additional mass. Another HCMV glycoprotein, gO (UL74), is 466 aa in length and is predicted to have a molecular mass of 54 kDa. However, the protein has an apparent mass of 125 kDa, and digestion with *N*-glycosidase produced a 65 kDa protein, suggesting that the *N*-linked glycosyl groups account for the additional 60 kDa (Huber and Compton, 1998). Based on this finding, Huber and Compton (1998) estimated that each N-linked glycosyl group adds 2-4 kDa of mass to a protein. Therefore, for N-glycosylation to account for the additional mass in the protein UL139, it would have to contain 15-30 *N*-linked glycosylation sites. This is considerably more than the three (G1), one (G2) and two (G5) predicted, and suggests that *O*-linked glycosyl groups and perhaps other forms of posttranslational modification may contribute to this additional mass.

Further experiments were not possible owing to time constraints, but the occurrence of glycosylation could be assessed through the use of deglycosylases and further immunoblotting. Other forms of post-translational modification, such as phosphorylation, may also contribute to the large mass of the UL139 protein. In addition, the RADs could be used to investigate UL139 localisation by indirect immunofluorescence, and to generate polyclonal antibodies against the UL139 protein.

5 Assessment of the genetic content of an AD169 variant that contains the U_L/*b'* region

5.1 Introduction

The following is a brief history of HCMV strain AD169. The virus was isolated in the USA at the National Institutes of Health (NIH) from the adenoids of an HCMV-positive seven year-old girl and passaged 14 times in human embryonic fibroblast cells (HEFs) to produce a stock named NIH 76559 (Rowe *et al.,* 1956). AD169 has subsequently been distributed throughout the world and is one of the most commonly used laboratory strains of HCMV. As more than 50 years have passed since AD169 was first isolated, the precise passage history of individual stocks is unclear, and numerous stocks exist. The two lineages that set the context of the present work are AD169*var*UK, which was developed by passage of NIH 76559 in the UK, and AD169*var*ATCC, which is a reference stock available from the American Type Culture Collection (ATCC).

In the UK, NIH 76559 was passaged ten times in HFFFs, four times in HEFs, eighteen times in human embryonic lung fibroblast cells (HELFs) and eight times in MRC-5 foetal fibroblasts cells. The resulting stock was then used to make batches of an HCMV vaccine by passaging 16-24 times in MRC-5 cells (Elek and Stern, 1974). A vaccine stock was plaque purified twice and used to produce a set of plasmid clones (Oram *et al.,* 1982). These clones were then used to sequence the genome of AD169 (AD169*var*UK), which was deposited in the public databases under accession number X17403 (Chee *et al.,* 1990).

The AD169*var*UK genome was characterised as being 229,354 bp in length, with U_L 166,972 bp, U_S 35,418 bp, R_L (TR_L and IR_L) 11,247 bp and R_S (TR_S and IR_S) 2,524 bp. The *a* sequence is 578 bp. Comparison with other HCMV strains revealed that the AD169*var*UK sequence contains frameshifts in three genes, namely RL5A, RL13 and UL131A (Akter *et al.,* 2003; Davison *et al.,* 2003, 2003a; Yu *et al.,* 2002). As well as these frameshifts, comparison with strain Toledo showed that AD169*var*UK has a large deletion of a region at the right end of U_L (15 kbp), termed U_L/*b'*, which encodes 19 ORFs (UL148-UL150) (Cha *et al.,* 1996; Davison *et al.,* 2003a; Dolan *et al.,* 2004). Moreover, this region has been replaced by an inverted duplication of a sequence of 11 kbp from near the left end of the genome that contains RL1-RL12 and part of RL13. This duplication results in a substantial expansion of R_L in comparison with low passage strains (Davison *et al.*, 2003a; Dolan *et al.*, 2004; Prichard *et al.*, 2001). It is probable that these mutations are a result of selective pressure placed on the virus during passage in cell culture.

In the USA, AD169 was deposited with the ATCC by W. A. Chappell (Centers for Disease Control and Prevention, Atlanta) and is now designated VR-538. The ATCC is unable to reveal the date on which it was deposited, but according to the literature this variant has been available from the ATCC since 1973 (Smith and de Harven, 1973). AD169*var*ATCC was cloned as a bacterial artificial chromosome (BAC) by insertion of a BAC vector immediately after US28 with no deletion of viral sequences (Yu *et al.*, 2002). From this clone, AD169*var*ATCC was sequenced (accession number AC146999) and used for further studies (Murphy *et al.*, 2003). AD169*var*ATCC has the same deletion of U_L/*b'* that characterises AD169varUK,as well as the duplicative expansion of R_L and the frameshifts in RL5A, RL13 and UL131A. Therefore these mutations occurred prior to separation of the lineages that led to AD169*var*UK and AD169*var*ATCC. Both variants also contain a single point mutation in UL36, which causes substitution of a C by an R residue and results in inactivation of the encoded inhibitor of apoptosis (Skaletskaya *et al.*, 2001).

Despite these shared mutations, AD169*var*ATCC differs from AD169*var*UK in replication efficiency (Brown *et al.*, 1995). In fact, various stocks of AD169*var*UK exist, and some contain an additional 929 bp of sequence within U_L that results in a decrease in the length of UL42 and an increase in the length of UL43 (Dargan *et al.*, 1997). This additional sequence was also found in AD169*var*ATCC and may contribute to the difference in replication efficiency (Mocarski *et al.*, 1997). Direct comparison of the AD169*var*UK genome with the AD169*var*ATCC genome by restriction endonuclease digestion revealed no other differences (Mocarski *et al.*, 1997), but sequence comparison indicates approximately 50 additional nucleotide substitutions plus approximately ten small insertions or deletions.

Unpublished work by Prof. N. Lurain (University of Chicago) using an AD169 virus stock, termed AD169*var*UC, which she received from Prof. Ken Thompson (University of Chicago) who in turn had obtained it from Prof. Marc Beem (University of Chicago) in 1981, indicated that this variant contains some of the genes in U_L/*b'* (N. Lurain, personal communication). Similarly, DNA microarray studies performed on AD169*var*UC in the laboratory of Prof. P. Ghazal (Laboratory of Clinical and Molecular Virology, The Scottish Centre for Genomic Technology and Informatics, University of Edinburgh) revealed evidence for the expression of genes in this region. These findings suggest that this variant may be an alternative passage of AD169 that contains part or all of U_L/*b'*. AD169*var*UC is the focus of this chapter.

5.2 Validation of the identity of AD169*var*UC

In initial experiments, several ORFs in AD169*var*UC were amplified by PCR and sequenced. RL5A, RL13 and UL131A were selected as they contain mutations in AD169*var*UK and AD169*var*ATCC (Section 5.1). UL11 and UL73 were chosen as they are hypervariable, falling into three and seven genotypes repectively (Hitomi *et al.,* 1997; Pignatelli *et al.,* 2003). The primers that were used to amplify and sequence these genes are listed in Table 2.3 and highlighted in blue as is the PCR product obtained. PCR products were either sequenced directly or cloned into pGemT with at least two clones being sequenced. The sequences were assembled and compared to those published for AD169*var*UK and AD169*var*ATCC. A summary of the comparison is shown in Table 5.1.

UL11, UL73, UL131A and RL13 are each identical in AD169*var*UC, AD169*var*UK and AD169*var*ATCC. RL5A in AD169*var*UC has a single nucleotide difference from the other two variants with RL5A. To determine whether AD169*var*UC contains U_L/*b'*, attempts were made to amplify and sequence three genes in this region (UL139, UL146 and UL148) using primers detailed in Table 2.3. All were detected (Table 5.1). UL139 and UL146 were sequenced as part of the genotyping study described in Chapter 3, the former falling into genotype G7 and the latter into genotype G9 (Table 3.1). The part of UL148 that is present in AD169*var*UK and AD169*var*ATCC is identical in AD169*var*UC (Table 5.1).

Table 5.1: Initial comparison of AD169*var*UC, AD169*var*ATCC and AD169*var*UK

ORF	*Length of ORF (bp)*[a]	*Present in* UK/ATCC[b]	*Nucleotide differences*[C]	*Number of Clones*[d]
RL5A	271	Yes	1	0
RL13	444	Yes	0	2 (0)
UL11	828	Yes	0	0
UL73	417	Yes	0	3
UL131A	391	Yes	0	5
UL139	441	No	NA	0
UL146	357	No	NA	0
UL148	951	Yes	0	3

[a] Includes the stop codon.
[b] UK, AD169*var*UK; ATCC, AD169*var*ATCC. The ORFs listed are identical in sequence in these two variants. Only 402 bp at 3'-end of UL148 is shared by the three variants.
[c] Between AD169*var*UC and AD169*var*UK/AD169*var*ATCC. NA, not applicable.
[d] Zero (0) indicates that the PCR product was sequenced directly.

5.3 Sequence of the U$_L$/*b'* region in AD169*var*UC

The U$_L$/*b'* region, plus flanking sequences, in AD169*var*UC was amplified as eight fragments by PCR (Table 2.3). The products were sequenced directly or as plasmid clones. The primers used for PCR and sequencing are listed in Table 2.3. Figure 5.1 shows the annotated sequence. AD169*var*UC contains the remainder of gene UL148 and the intact genes UL147A, UL147, UL146, UL145, UL139, UL138, UL136, UL135, UL133, UL148A, UL148B, UL148C, UL148D and UL150, all of which are absent from AD169*var*UK and AD169*var*ATCC. In addition, the contiguous sequence that was obtained contains, to the right, the internal inverted repeat sequences (IR$_L$ and IR$_S$, which are described alternatively as *b'-a'-c'*), the latter of which contains part of gene IRS1, and, to the left, genes that are also present in AD169*var*UK and AD169*var*ATCC, namely UL128, UL130, UL131A, UL132 (none of which are included in Figure 5.1) and the 5'-part of UL148. A non-contiguous sequence to the left of U$_L$/*b'* containing UL121 in its entirety and parts of UL122 and UL123 was also sequenced. The results of comparisons of these sequences with those of AD169*var*UK and AD169*var*ATCC are shown in Table 5.2.

The sequences of UL121, UL122, UL123, UL128, UL130, UL131A, UL132 and the shared part of UL148 are identical in the three variants. Unlike the other variants, AD169*var*UC contains most, but not all, of U$_L$/*b'*. A 3.2 kbp sequence is absent, resulting in deletion of the entire UL141 and UL142 ORFs plus 148 bp at the 5'-end of the UL144 ORF and 27 bp at the 3'-end of the UL140 ORF. This deletion is highlighted in Figure 5.1. It results in a framshift near the 3'-end of the UL140 ORF, with the C-terminal 8 aa replaced by 71 aa in an alternative reading frame. The alternative C terminus is highlighted in grey in Figure 5.1.

Table 5.2: Further comparison of AD169*var*UC with AD169*var*ATCC and AD169*var*UK

PCR product name	*ORFs present*	*Sequence obtained (bp)*	*UC/UK*	*UC/ATCC*	*UK/ ATCC*	*Length of ORF (bp)^{a}*	*Number of Clones^{b}*
RS1/RL1	TRS1	2711^{c}	8	23	17	609/2367	0
	RL1					817/936	
UL122	UL121	3165	0	0	0	543/543	5
	UL122^{e}					1487/1743	
	UL123^{e}					818/1476	
UL132	UL128	2851	0	0	0	516/516	5
	UL130		0	0	0	645/645	
	UL131A		0	0	0	391/391	
	UL132		0	0	0	813/813	
UL146	UL132	2853	0	0	0	813/813	3
	UL148		0	0	0	951/951	
	UL147A		NA	NA	NA	228/228	
	UL147		NA	NA	NA	480/480	
	UL146		NA	NA	NA	357/357	
Four smaller products	UL145	887	NA	NA	NA	393/393	0
	UL140	1197	NA	NA	NA	765/765	0
	UL139	890	NA	NA	NA	441/441	
	UL138	1240				510/510	
UL136	UL138	1592	NA	NA	NA	510/510	0
	UL136		NA	NA	NA	723/723	
	UL135		NA	NA	NA	987/987	
UL133	UL133	2073	NA	NA	NA	774/774	0
UL150	UL148A	2892	NA	NA	NA	240/240	0
	UL148B		NA	NA	NA	243/243	
	UL148C		NA	NA	NA	234/234	
	UL148D		NA	NA	NA	195/195	
	UL150		NA	NA	NA	1917/1917	

^{a}Includes the stop codon.
^{b}A zero (0) indicates that the PCR product was sequenced directly.
^{c}This comprised part of R_S sequence and the entire R_L, plus 878 bp from the left end of U_L.
Note- The PCR products UL132 and UL146, and UL136, UL133 and UL150 overlapped.

The sequences of UL147A, UL147, UL146 and UL145 (to the left of the deletion), UL139 (to the right of the deletion), and the residual parts of the UL144 and UL140 ORFS were compared with sequences from other HCMV strains in Genbank using BLAST.

```
CTACCGACGCCGCGACACCAGGTAGGTTATCAAAACGCGAGCCCATATCGCCGCCATCATTGTAATCAGCAATGTGTTGAGGTACTGCACGATGAATCTG   100
 -  R  R  R  S  V  L  Y  T  I  L  V  R  A  W  I  A  A  M  M  T  I  L  L  T  N  L  Y  Q  V  I  F  R

TCTAGTGACACCAGCCAACCCTCTGCTTTTGCGGGCAAGCGCGCTTTCGGTGACAGGGTGTATCGTACGTAGCCGCGGGTCAGGCGCGCGTTGTAGCGGT   200
D  L  S  V  L  W  G  E  A  K  A  P  L  R  A  K  P  S  L  T  Y  R  V  Y  G  R  T  L  R  A  N  Y  R  Y

ACACGCAGAAATCTATCCACAGGCCAACGCCCGGCTGTAGCTTCGGATGGTGGATAATAGCGCGGTGACGTACGCCGCGTGGCTTTAGAATCTCCACCTG   300
  V  C  F  D  I  W  L  G  V  G  P  Q  L  K  P  H  H  I  I  A  R  H  R  V  G  R  P  K  L  I  E  V  Q

TAAGGCCATCTCCTCCAGGTAGTGGGTCTGACTGCGACGCAGCGTCCAGTTCATGTAAAAGTCGGTCTCGCCGTGTCCGGCCACGAAGAGGCTGCTTACT   400
 L  A  M  E  E  L  Y  H  T  Q  S  R  R  L  T  W  N  M  Y  F  D  T  E  G  H  G  A  V  F  L  S  S  V

AAATCGGGCGCCAGAGCTAGGTCAGGCGTATCAAATTCCACTGCCAGGCGACCTGATTCTAACGGTTCCACGATCCGGGAGAGCGTTTCTAGATATAGAG   500
L  D  P  A  L  A  L  D  P  T  D  F  E  V  A  L  R  G  S  E  L  P  E  V  I  R  S  L  T  E  L  Y  L  A

CAAAGCGTACCACGTCTACCTGCGGTGTAAAAAACTGCTGTGGGCGTTCACCGTCGTTGACCACGTAGGCCACGTAGAGGCCAACATTTTCCACCACGGG   600
  F  R  V  V  D  V  Q  P  T  F  F  Q  Q  P  R  E  G  D  N  V  V  Y  A  V  Y  L  G  V  N  E  V  V  P

TTCTAGCTGCAGGCGGCACGTAAAGCTTAGAAACGACGGCTGTACGGTTTGGTTCCCGTGAAGCTGAAGCGTCACTTCCTTGCCGGGGCTCACCGTGCTG   700
 E  L  Q  L  R  C  T  F  S  L  F  S  P  Q  V  T  Q  N  G  H  L  Q  L  T  V  E  K  G  P  S  V  T  S

TAACGCCGCACCGAGTCGGTCATCTGCTCCAGATCGGTAGACCAGAAGGGTGTGCAATGCATACTGTCCCAGTCGCGACACGCAGCCCAGCCTAGCTCGG   800
Y  R  R  V  S  D  T  M  Q  E  L  D  T  S  W  F  P  T  C  H  M  S  D  W  D  R  C  A  A  W  G  L  E  T

TGAAGGGTCGACGCACACCCGAGAAAGTGTGCTTGAAGACCAGGGGGTCGCCTCGGTAGCTCAGTAGCCGAACATGCACATAGTCGCGGCTAGCGTTGAC   900
  F  P  R  R  V  G  S  F  T  H  K  F  V  L  P  D  G  R  Y  S  L  L  R  V  H  V  Y  D  R  S  A  N  V

AGACGGCCCGTGGAGGGCCAGCAGGACAAGCGTGAACAGCAAGCGCAACATGCTGCGCGGGTTAGGAAATGCGGCGTGCCGGCCACCGCCCGACTCATAA  1000
 S  P  G  H  L  A  L  L  V  L  T  F  L  L  R  L  M  UL148

ACGCTACCAGCATGACGTTTCAGATCACACAAGTGACGAGGAGCGTACCGCAAATCACTAGGGAAAAGGCCAGCAAAGCCCGATAGTCTTGCTCTTCGCG  1100
                    -  I  V  C  T  V  L  L  T  G  C  I  V  L  S  F  A  L  L  A  R  Y  D  Q  E  E  R

AACGATCTCGTCCGGTTCCTCGCAATCTTCGTGGTCCACAGAAGACGAGGAGCAGGAGTTTTCGTTAATCTCTGCGAGGATACTAGTGCTATACCACACC  1200
 V  I  E  D  P  E  E  C  D  E  H  D  V  S  S  S  S  C  S  N  E  N  I  E  A  L  I  S  T  S  Y  W  V

AGAGCGCTCAGTGTGCCCAGAGCTACCGCACGGTAAAATAGGGACATGATCACCAGCGCAGTCTAAAGTAGTGGTAATTCAGTTTCTTGGCGTATTTCCA  1300
L  A  S  L  T  G  L  A  V  A  R  Y  F  L  S  M  UL147A
                                                   -  W  R  L  R  F  Y  H  Y  N  L  K  K  A  Y  K  W

GAGAAAGGCTTTGTAGGCCGTAGGGACTGGCCAGGCACCGAACTCAATATTTGTAGACACTACGTCGTAAATGCGTTGTTCCTCGTCTAAGACTAACCGA  1400
 L  F  A  K  Y  A  T  P  V  P  W  A  G  F  E  I  N  T  S  V  V  D  Y  I  R  Q  E  E  D  L  V  L  R

AAAAATAGCCGATTGATGTGACGACGCACGGCTTGCGCGTTAGGATTGAGACACTTGGTGCCCTTGTCCTTTAAAATAGCCAGCACTTCCTGACGATTGC  1500
F  F  L  R  N  I  H  R  R  V  A  Q  A  N  P  N  L  C  K  T  G  K  D  K  L  I  A  L  V  E  Q  R  N  C

AGCTTTCGCTCGCTGCGATTGGCTTAAGTAATTCGGTTCCTATTGGCAGGGTATTCAACAGAATTTGGTTGTTACAACGACAGCGCCGGTCGTAATCTTC  1600
  S  E  S  A  A  I  P  K  L  L  E  T  G  I  P  L  T  N  L  L  I  Q  N  N  C  R  C  R  R  D  Y  D  E

CAGCTCTAGAAGATGGACAACTGGGGGACACACGGCAAATAACATATATGCGGTCAAAGACAGGTGTCGTACCGATAAAAGTTTTATATGCGAATTCGAA  1700
 L  E  L  L  H  V  V  P  P  C  V  A  F  L  M  Y  A  T  L  S  L  H  R  V  S  L  L  K  I  H  S  N  S

ATCGGATGATGTAACCATGTTAACATCATATCGAAAACATATTGCGTTATCGTTTCTTGTAAAAATTTTATCAACTATACACATATTACTGATTCGTTAA  1800
I  P  H  H  L  W  T  L  M  UL147                                                                  -

AATTTTAGTTTCCAAGGCGGACGTCCGTTACTAGCAGTTCTTTCCTCTACGTGCGGTCCACCACCACCTTTTGTTCTTAATAACACTTTGTGCCATGTGT  1900
F  K  L  K  W  P  P  R  G  N  S  A  T  R  E  E  V  H  P  G  G  G  G  K  T  R  L  L  V  K  H  W  T  N

TACTTGATTTGCCATGAAGCCATTTAGATAACACATGATCAGGAGACAAACACACAGGTTTACCTTTAGGAGGCAATAAAAAATGTTGCGGCTTTTCACA  2000
  S  S  K  G  H  L  W  K  S  L  V  H  D  P  S  L  C  V  P  K  G  K  P  P  L  L  F  H  Q  P  K  E  C

TTTAGGAGGATCTGGGGGATTATATCCAATCCAAAAAAAGCCACCTATTGGATAGCTTAAACCATTACTACCACAAGGGCAACGTAACTCCACACTCTCA  2100
 K  P  P  D  P  P  N  Y  G  I  W  F  F  G  G  I  P  Y  S  L  G  N  S  G  C  P  C  R  L  E  V  S  E

ACTTTATAGTACAACGCGATCAAAAGACCAAACAGACTAAAAATAAATCGCATAATTTTATTAGCTACGTCACTATCAGTAATTCGTAATATCCGGTATT  2200
V  K  Y  Y  L  A  I  L  L  G  F  L  S  F  I  F  R  M  UL146

CCCGGAAAATCACTCAAAACTGCGTCCATGACACATCGATTCCCGATAACTACCTCCCTTTGAAATCGGATCCCCCCACATACCAATCAATCACACAACA  2300

CACAGGTTTAAAAATCGATCACACGTCAATTAGGTTTCAAAATCGATACTGTTTATTATCAGGAATCTAGACTAATTCTACAATGACAGCTCTGAATTTC  2400

TCTCTCGTCTTTCTTGTCAGGTTCTCATCATCAATCTTCACTTCCACCCATCGAGGAGTCATCGTCGCTCCAAAACCCTTTGGGGTCGCTGGTTGGAAAA  2500
                               -  D  E  S  G  G  M  S  S  D  D  D  S  W  F  G  K  P  D  S  T  P  F

GTCTCTGACACGATCCAGGCACCCCGTACCCAGTCCGACTGATCTAGCTTACGGAGCATCTCAACAGGCATGAGCTGCAGGGCCACGGCTGTCACGGCAC  2600
T  E  S  V  I  W  A  G  R  V  W  D  S  Q  D  L  K  R  L  M  E  V  P  M  L  Q  L  A  V  A  T  V  A  S

TGTATCGATGTAACACTAGAGACTTTCTTTGCGATGTAGCCATCAACACGGCATATGCTCCATAGTTCGCGTGATACGACGCATGATGGGTTAAACGTTC  2700
  Y  R  H  L  V  L  S  K  R  Q  S  T  A  M  L  V  A  Y  A  G  Y  N  A  H  Y  S  A  H  H  T  L  R  E

CCATCCGGCAGTGCCGTCTCGGGTCCGTGCACACAACAGCTGCACGGCGTTATGATGCTTAAAATTAACCATAACGCTGGGGCTACTGATAAAGGAGTAG  2800
 W  G  A  T  G  D  R  T  R  A  C  L  L  Q  V  A  N  H  H  K  F  N  V  M  V  S  P  S  S  I  F  S  Y

TAATGAGCCAGGACGCCGTACATCGAAGGCAACAAGAAAGAGTGACAGCACGATAGCACCGGGCTCTTATGTAGGCGACAGCTTATTTTTCCTGACGTCG  2900
Y  H  A  L  V  G  Y  M  UL145

GCAAAAAGTACCTAAATTCCCCACAGATATTCAGACACGGTTCCGTAAAGTGCTTCTTTTTTTAGTGCAGGAATTGGAAAAAAATAATAAAAAATATGAAC  3000

AGCTCATCTGTAATTATCTGTGTGACTTCATCGTACCGTGATGTAAAAACAACAACAGGAAGCCTACAGGGTGCGGTAGAAAATTTTGCCGATTGAGCAA  3100
```

Figure 5.1 The nucleotide sequence of the right end of U_L in AD169*var*UC
(Continued overleaf)

```
CACTGTTGGCATCTCTCACTCCGATAGGCGGCTATAAGATAGAGAATTAAAAGTATGATACCCACGAGAAAGATGAAGAGGGACAACCAGGCTAGAGTAT  3200
  S  N  A  D  R  V  G  I  P  P  -  L  I  S  F  -  F  Y  S  V  W  S  F  S  S  S  P  C  G  P  -  L  I

GACGACCACTTTTCCCTTGTTTGACGGTTACATGTGCGGTATGATTTTGTCGTTGCTTGTGATGTTGGACGCCTGGAATGGAAAACGACGTATAATTCTT  3300
 V  V  V  K  G  K  N  S  P  -  M  H  P  I  I  K  D  N  S  T  I  N  S  A  Q  F  P  F  R  R  I  I  R

AGATGCGCATACGGTGTTATTAGTGGAAGTGCAGTTACGAATCGTAACCTCAGTGTCATTACACTCAGTGCAATTGGTACAATTGTAAAGCCCTGATACA  3400
L  H  A  Y  P  T  I  L  P  L  A  T  V  F  R  L  R  L  T  M  V  S  L  A  I  P  V  I  T  F  G  Q  Y  M

TACGTACCGTTAGGGCAAAGTGTACATGTTGTACTCGTATATTGTCGTGGCTGTGGCGCAGCGCGTTTTGTACAGCGCGTTTCACCGCTTTCTGCATGGC  3500
  R  V  T  L  A  F  H  V  H  Q  V  R  I  N  D  H  S  H  R  L  A  N  Q  V  A  R  K  V  A  K  Q  M  A

CGCTACCACATCGGGCGGGAGTGGCTCCGGCGGAAGCTCGATGAGCAGTTGCTGCGAATCTCGGCGCTCGGTGTCCGCCGTTTCGTCGGACGTGGCGTAA  3600
 A  V  V  D  P  P  L  P  E  P  P  L  E  I  L  L  Q  Q  S  D  R  R  E  T  D  A  T  E  D  S  T  A  Y

AAAACCGAGGTGGTCGCCCAGTCGTCCACGCTGTCGACGGCTTCTGTTAGTGCCGGGTTGTCAAAACCGCCATCGGACGCGGGTGATAAAAGAACGTACG  3700
F  V  S  T  T  A  W  D  D  V  S  D  V  A  E  T  L  A  P  N  D  F  G  G  D  S  A  P  S  L  L  V  Y  S

ATGACACGCTGTTAGTACGACTCTCGTCGTCGCTTTGGGAACGACGTGATGGACGACGGTAGATGACCTCGTCTTGCCACGCGTCGAAGCGGTCGCAGCA  3800
  S  V  S  N  T  R  S  E  D  D  S  Q  S  R  R  S  P  R  R  Y  I  V  E  D  Q  W  A  D  F  R  D  C  C

GCGCTGGATCCAAGCGCAGCGGAGCAGCTTACGGAACACGTCGTTGTTCCAAAAGTAGAGCATAAAAAGAAAGAAAAGTAGCGTAACGATGAAGCCGAAA  3900
 R  Q  I  W  A  C  R  L  L  K  R  F  V  D  N  N  W  F  Y  L  M  F  L  F  F  L  L  T  V  I  F  G  F

ACGACGAGGGTCGGCAGGGCACTGCCGCCGCTGCCGTTTTTTGCGTCGTGCGGGTGCACGGTGGTAGTGGCGTTAGTCTGAGCGGGGGTCATGACAAGTC  4000
V  V  L  T  P  L  A  S  G  G  S  G  N  K  A  D  H  S  H  V  T  T  T  A  N  T  Q  A  P  T  M  UL140

TGAAGAGATGAGAGCGCGGGTGCTCATCAGGAACAGTTGAGGTCTCTCCCTACCGAAGCCTTAGCCTCTACGGTGTTTTATGATGAACGTGTATACGAAC  4100

GTCATTGTGAAAGTGACGTCTCAGGCCTTCCGAAACCGCGTTAGGTTCAACGTGGGTTTCGGTTTAGCCTGCGTCACCGAGGCGGAGGTGGAAATGAGCC  4200
                                                                          -  R  P  P  P  P  F  S  G

GTCCTGTGGGGGAGTGTACGACCCTGTAGTGCCCATGGGTAACGTTGCGTCGGAAGAAGTGAATGCGGCATTGGTGTACGCGTGGGTTGTTTTGCTCTCT  4300
 D  Q  P  P  T  Y  S  G  T  T  G  M  P  L  T  A  D  S  S  T  F  A  A  N  T  Y  A  H  T  T  K  S  E

GACTCGGAGGAGTTGCCGCAGCAGCTGCAGATTTTACGTACTAGCCAAAAGCAGCAAAAGCAGCAGGTAAATAAGAGAAGGAGTCCAGATAATGTCCAGT  4400
S  E  S  S  N  G  C  C  S  C  I  K  R  V  L  W  F  C  C  F  C  C  T  F  L  L  L  L  G  S  L  T  W  D

CGCTAGCGGCAAACAGCGCAAGTTGCGCGACTGTCCAATTACTGCCACCAAAGGCTTGAACACATTCAATGGTTGGTGTTGCAGTGCTGTTATTGCTACT  4500
  S  A  A  F  L  A  L  Q  A  V  T  W  N  S  G  G  F  A  Q  V  C  E  I  T  P  T  A  T  S  N  N  S  S

AATGGACGAAGAAGACGAAGACGACGAAGTAGCTTGACTGGAATTAGAGCTGGTACCTGTAGTGGTTTCACTCGCCGATGCGGCAAGTGCAAATAAAACT  4600
 I  S  S  S  S  S  S  S  S  T  A  Q  S  S  N  S  S  T  G  T  T  T  E  S  A  S  A  A  L  A  F  L  V

AATATCCACAGCATGTTCGTTACTATATAATTGATATACGAACCCGTTTGTCGTAACAATCAGCGTTATACACGCTGTATCGGCATCGTTTTACTGGAAA  4700
L  I  W  L  M  UL139

GTTTATCGTAATGTAACCCGCGTTGTGTACATTCGTACTGACAGGGAACTCCCGGTGATGTGCACATTATACTCTTTCATTCTGGGGTTTCCCAATGACG  4800

TAAAAATTTCCACTACACAATAAAATTACGGACTCATGTGAAAAGTGTGCTTTTTATTAACAGAGCAGAGGGTTTACAGTAGATATATGTTTGCCAGGGC  4900

CACCGTTTTCTAACACCGATCACCGCCACCATTACCACCCGTTGAACTCCACACCCGGGAGCCGCCTGATCGCCAGGGACTCCTCACCGTCCATCGTCCG  5000

AACAAGCTCCCGCCACCGATGCTGCCACCATCACCGAGAGAAAAAACCGCTTGCTGCAGATACGCTTGGGCTCGCCTCCGTGCGGACGCCGTTTCGTGCA  5100

GACGCTGAGTAGATCGAGCAGAGAATGTCAAAGCGACATTATCGCGATCCGCTCCCCTCTTTTTTCTTTTTCTCATTCACGTGTACTCTTGATGATAATG  5200
                                                                             -  T  Y  E  Q  H  Y  H

TACCATGGCTACGGTGGTGAACTGCGTCGCGGATCCCGTCACGGGTTTCAACAGATCGACGTCGGTCAGCGGCGCCGTCACCGCCATGTCCGGCGGAGGC  5300
 V  M  A  V  T  T  F  Q  T  A  S  G  T  V  P  K  L  L  D  V  D  T  L  P  A  T  V  A  M  D  P  P  P

ACGCTGTTTCTCTGGCTAGCGACGTGGACCGACGACGAAGACGATGAACCCGCGCGGCGGTCTGTTATCCGCGACGACGCGTAGCTGCACTGGGAAGACA  5400
V  S  N  R  Q  S  A  V  H  V  S  S  S  S  S  S  G  A  R  R  D  T  I  R  S  S  A  Y  S  C  Q  S  S  V

CTTCCTCCCAACGGACCAAGATCTCGTCGGGCCGTTCGGAGAAACGGTATCGTCTGTCCGACTCCCGCCGTACGGCGCCGAGGCCCAGAGACGACAGGTC  5500
  E  E  W  R  V  L  I  E  D  P  R  E  S  F  R  Y  R  R  D  S  E  R  R  V  A  G  L  G  L  S  S  L  D

CGCGAACCGGCGCTCGTACTCCCCGTACAGCTCGCAACAGCGGATCAGCCAGCGGTAGCTCAGAAACATGCGCACTAGTTTGAAGGTGTCGTGCCAGTGG  5600
 A  F  R  R  E  Y  E  G  Y  L  E  C  C  R  I  L  W  R  Y  S  L  F  M  R  V  L  K  F  T  D  H  W  H

TAAGCCAGATAGCAGAGGATGGCCACGATCAGCACGAGCATCACGCCGATGATGGGTAACCCGACATTCAGCGGCAGATCGTCCATGGTGACCGTCCTCT  5700
Y  A  L  Y  C  L  I  A  V  I  L  V  L  M  V  G  I  I  P  L  G  V  N  L  P  L  D  D  M  UL138

GTCCGGATCTACGTCCCAGTCTCTCTCTTTTGTACAGCACTCGCGCGGGAACGGCCCCCTCAACCCTCTTACGTAGCGGGAGATACGGCGTTCTCCCGCG  5800
                                                                     -  T  A  P  S  V  A  N  E  R  P

GGCCACTTACTTGCACGGTCGCTTGAACGGCGGCTTGGACCGCCACATGCACCGCATCCATCCATTCCGGCAGCAGCGCGTTCGGCGACGTCGTACGAGT  5900
  G  S  V  Q  V  T  A  Q  V  A  A  Q  V  A  V  H  V  A  D  M  W  E  P  L  L  A  N  P  S  T  T  R  T
```

Figure 5.1 The nucleotide sequence of the right end of U$_L$ in AD169*var*UC

(Continued overleaf)

```
GCACATCCATGGCTCGCCGTCTCCTTCTCTGCCGCTCGTGGTGCCGACGGCACTTCTCGGGATAATGACAGCCGCAAAATAGATCGTGGAGCATGTCTCG  6200
  V  D  M  A  R  R  R  R  R  Q  R  E  H  H  R  R  C  K  E  P  Y  H  C  G  C  F  L  D  H  L  M  D  R

CCAACTGTCCTGGTGGTAATATCTTAAGTACGCGATGAGCGCGCCGATGGCCATAATCATAAGCGTAAGCAAAACGGCACAGATAACGTGAAACACCGCG  6300
 W  S  D  Q  H  Y  Y  R  L  Y  A  I  L  A  G  I  A  M  I  M  L  T  L  L  V  A  C  I  V  H  F  V  A

GTCATCCAAGTCGGGCGGCGTCGGGGACGCGGTGGGTCGGTTTCTCTTACGCCGGCGTCACTCAGCCACCACACCCGTAGCCGACATTCCCAGAACCGGT  6400
T  M  W  T  P  R  R  R  P  R  P  P  D  T  E  R  V  G  A  D  S  L  W  W  V  R  L  R  C  E  W  F  R  H

GAATGCGACTCAGGGCCTTTCGACGCCGCCATTTATTTCCAACGTCCAAGTCCCACGTCATTTCTGGCATCTCCACGCCCTTGACTGACATACTCTCTTT  6500
  I  R  S  L  A  K  R  R  R  W  K  N  G  V  D  L  D  W  T  M  E  P  M  E  V  G  K  V  S  M  UL136


CTCTCTCTTAGCTGCGGTGAAAAAGAGGGAAGGCGTGTGCTGCTATACAACTGTACAACGGACGCGCTCGCTCTTTCGGTCTCAGGTCATCTGCATTGAC  6600
                                                                                  -  T  M  Q  M  S
TCGGCGTCCTTCATGACGCTCTGCACCGCCTTTTCCAAGAGTTCCTCGATGTCCGACCATCGAGGAGGCGGGGCTAACTCGGAAACCGACACGATAGGCA  6700
E  A  D  K  M  V  S  Q  V  A  K  E  L  L  E  E  I  D  S  W  R  P  P  P  A  L  E  S  V  S  V  I  P  L

GCGTGGTCGGCTCCGTCGGCGTGCGGGGTCGGGGACAGGGACACGAGAGTCCCACCTTCGAGAGATTCTCCAGCCCGACGGTGCGCGGCAGTCTCGGATT  6800
  T  T  P  E  T  P  T  R  P  R  P  C  P  C  S  L  G  V  K  S  L  N  E  L  G  V  T  R  P  L  R  P  N

CCGCGGTGGCTTTTGTGGCGTCGGCGTTTTCGGGAAGGGCCTGGGCGTCACCGGCGGTGTCCAGCCGACCGGCTTGGGTTTCGTGGGCGGCGGTGTTTTC  6900
 R  P  P  K  Q  P  T  P  T  K  P  F  P  R  P  T  V  P  P  T  W  G  V  P  K  P  K  T  P  P  P  T  K

TTGGTGGGCGGCGTGCTCAGGTTCTTACGCGGCGCGGGTATCGGCGTCGGGGGCCTGTGCGACGACAGCCGCGTGGTGGGGGCCCGGACCGGCGGCGTAG  7000
K  T  P  P  T  S  L  N  K  R  P  A  P  I  P  T  P  P  R  H  S  S  L  R  T  T  P  A  R  V  P  P  T  P

GCGGCCGCTTCTTGCGCCCGGGCGGCGGAGGTGGCTTCCAGGATGGCGGCGGCTGATGCAGTACCGTGTCGACGCTGGCCGAGGACGACAAAGAGCTCGA  7100
  P  R  K  K  R  G  P  P  P  P  P  K  W  S  P  P  P  Q  H  L  V  T  D  V  S  A  S  S  S  L  S  S  S

CGAGGAGCAATGCGACGGAGATCGGCCGATGCTGGTCGGCGTTCCCGGCGTGGATACGTCGGGGATCTCGAATCGCGCCGGAGGAAACTCGGGTTTATCT  7200
 S  S  C  H  S  P  S  R  G  I  S  T  P  T  G  P  T  S  V  D  P  I  E  F  R  A  P  P  F  E  P  K  D

ATCGGCAGACCATCCTCTCCTATGTAGAGCGACGTACACCGCGGCACCTGCGGCGTCGGCGGGTGGGTGGCCACCCGCATGAGCCCCAGTTCCAGATCCA  7300
I  P  L  G  D  E  G  I  Y  L  S  T  C  R  P  V  Q  P  T  P  P  H  T  A  V  R  M  L  G  L  E  L  D  L

GCGGCTCGACGACGTCTTCTTTCGGAATTCGATAGCAGCACGCGCAGACACCACGCTTATCAGAAGCAGCACCCGGGAGCCGGCCTCGCGACGAAGTCTC  7400
  P  E  V  V  D  E  K  P  I  R  Y  C  C  A  C  V  G  R  K  D  S  A  A  G  P  L  R  G  R  S  S  T  E

GTCGGATCGCTTGCGGCCTCGGCGCTGGGTAAATAAGGAAATGGCCAGGACCAGGGAAGCCAGTCCGGTACCGCCGAGAAGCCCGACGCCGAGCCATATC  7500
 D  S  R  K  R  G  R  R  Q  T  F  L  S  I  A  L  V  L  S  A  L  G  T  G  G  L  L  G  V  G  L  W  I

CACACCATGATCTTCTCTCCTGCTTGGAATCTCAAACTCCGTGTCGGGAAGGGCCGGTGTACGGACATTTATGCCTTGGATTTCTGGAAACGTCATTTTT  7600
W  V  M  UL135

TGGCAAGGAATGTGTTTATTGTCCAAACACTGAGGAAGGAGATGTGGGCCAAGTCGGAAAATTCCTTATCACACCGGGGGCGGGTTACGTTCCGGTCTGA  7700
                                                                                     -  T  G  T  Q

TGCTGCTGCTGTTGTTGTAGAGCCGCGGCCACGGCCGCCTGCACGGCAGCTTGTACCGCCTCGGCCACGCCGGGTGGCATCTGCGGCATGGCGGGGGGAG  7800
H  Q  Q  Q  Q  Q  L  A  A  A  V  A  A  Q  V  A  A  Q  V  A  E  A  V  G  P  P  M  Q  P  M  A  P  P  P

GCGCATCGGGCGGACCGCCGGGCATCGCCGTCGGCTGCGACGGTGGTTGTGAACTCACCGTCGGCTCGCACGGAGGTTTGTCCTTCGGTCTATCCTTCGG  7900
  A  D  P  P  G  G  P  M  A  T  P  Q  S  P  P  Q  S  S  V  T  P  E  C  P  P  K  D  K  P  R  D  K  P

TTTATCTTTCGCCCTACCTTTTTTCGGTTTGGGTTCCGATGTCGGTGCTGGCGGCTGCGGTGGGATGACGGGCTGGTGGAACTCCTCCGACGGCGGGGGG  8000
 K  D  K  A  R  G  K  K  P  K  P  E  S  T  P  A  P  P  Q  P  P  I  V  P  Q  H  F  E  E  S  P  P  P

ACGAACACCGTCGGCGCCGAAACCGGGGGACTCTCGACTATCTCGCAGATCACCCTGTCGGGATCGTCGCCGTGTCCGGGACGCCGTCGATGACCGTATT  8100
V  F  V  T  P  A  S  V  P  P  S  E  V  I  E  C  I  V  R  D  P  D  D  G  H  G  P  R  R  R  H  G  Y  Q

GGACCATGTCGTAAATCATCGTCTCCTTGTAACACGCTGAACAGCAGCGGCTGCAGGGACCCGAAATGCATTTGCAACTGCACTTACAGCTACAGCTGCA  8200
  V  M  D  Y  I  M  T  E  K  Y  C  A  S  C  C  R  S  C  P  G  S  I  C  K  C  S  C  K  C  S  C  S  C

GTAGCGCACCCATCGGCAAGTTAAAATGTCGATTATGGAATCTTTAAAGAATTCCCGGTAGCGGATGAGGTACGCGCAGAGGAAAATCATGAAAACCGAG  8300
 Y  R  V  W  R  C  T  L  I  D  I  I  S  D  K  F  F  E  R  Y  R  I  L  Y  A  C  L  F  I  M  F  V  S

CAGCCGACCACGGCTGCAATACCGGGTCCAGAAGAGAAATCCGATGACCATCCCGCCAAACACCAAATTCCCAAGGCCGCGCATGTTATCCAGGCCACAA  8400
C  G  V  V  A  A  I  G  P  G  S  S  F  D  S  S  W  G  A  L  C  W  I  G  L  A  A  C  T  I  W  A  V  I

TAATCGTGGGAACGCCCCATTGGCATTGCCACGAAGGATCGTGCACGTCGCAACCCATCGCTACTGCGTTCTCCCACAAACGCCATCGCACTATTTATCC  8500
  I  T  P  V  G  W  Q  C  Q  W  S  P  D  H  V  D  C  G  M UL133

CTACAGCGGCTGCCGAGTCACGTCCGCCGGCGCCCATCGGCCGCGGCGATCTCCTAGTAACACTCGTCCGACACTTCCACCATCTCCAGCTCGGCCGGCG  8600
                                                      -  Y  C  E  D  S  V  E  V  M  E  L  E  A  P  P

GTTCGGCATCCTCCACCAGCGGCGTCGTCTCATCTTTTCCGCAGCAGCGAACGCACACCTTCTCCAGGCAGAACGCCACCAGCTGCCGCCGAACGTACCA  8700
  E  A  D  E  V  L  P  T  T  E  D  K  G  C  C  R  V  C  V  K  E  L  C  F  A  V  L  Q  R  R  V  Y  W

CAGGTACACGTGCAGACCTGCGAACAGGACTACGGAGGTCATGACAACCACGACGCACACGGGAATCCAGGGATCGAGATTTTCGGAACCCATGGCTATC  8800
 L  Y  V  H  L  G  A  F  L  V  V  S  T  M  V  V  V  V  C  V  P  I  W  P  D  L  N  E  S  G  M  UL148A
```

Figure 5.1 The nucleotide sequence of the right end of U_L in AD169*var*UC

(Continued overleaf)

```
ATTACGGTGACCCCCATGACTAGACCCACGCAGATAGCCAGCCCCGCTAGCGTATCCAGCGCCATCCCGTTCGCTCCCGTCGTCGTCTCCTGAACAAAGC  9200
I  V  T  V  G  M  V  L  G  V  C  I  A  L  G  A  L  T  D  L  A  M  UL148B

                                              UL148C   M  L  T  P  A  V  F  P  A  V  L  Y  L  L  A
AACAACTCCACAGTCCCCGTTTTCAACCGTTTTTGTTTCCTTCTCCGCGACTAGATGTTAACGCCCGCGGTCTTTCCGGCCGTGCTCTACCTCCTGGCGC  9300

L  V  V  W  V  E  M  F  C  L  V  A  V  A  V  V  E  R  E  I  A  W  A  L  L  L  R  M  L  V  V  G  L  M
TTGTCGTCTGGGTTGAGATGTTCTGCCTCGTCGCCGTAGCCGTCGTCGAGCGCGAGATCGCCTGGGCGCTGCTGCTGCGGATGCTGGTCGTTGGCCTGAT  9400

  V  E  V  G  A  A  A  A  W  T  F  V  R  C  L  A  Y  Q  R  S  F  P  V  L  T  A  F  P  -
GGTGGAAGTCGGCGCCGCCGCCGCTTGGACCTTCGTGCGTTGTCTCGCCTATCAGCGCTCCTTCCCCGTGCTTACAGCCTTCCCCTGAAACCCACGTTAA  9500

CCGACCGTCCCGAAAACGCCGGTGTTAACACAGGAAAAAAAGAAATCACGCAGGAACCGCGCAGGAACCACGCGGAACATGGGACATTATCTGGAAATCC  9600

TGTTCAACGTCATCGTCTTCAGTCTGCTGCTCGGCGTCATGGTCAGTATCATCGCTTGGTACTTCACGTGAACCACCGTCGTCCCGGTTTAAAAACCATC  9700

ATCGACGGCCGTTATAAAGCCACCCGGACACGCGCCGCGGCACTTGCCTACGGCGCTGCTCCAGGGAAACTCCTCTTCCTCCTGCTCTTCCTCCTCCGCC  9800

                                                             UL148D   M  T  A  P  K  C  V  T  T  T
GCAGGGATCGTTTCCCTCGACTAGGGACCCGCCGAAGCAACTGCCGGAACAACCTGGAGGAGTCGCGGCATGACGGCGCCCAAGTGTGTCACCACCACTA  9900

T  Y  V  V  K  T  K  E  R  P  W  W  P  D  N  A  I  R  R  W  W  I  S  V  A  I  V  I  F  I  G  V  C  L
CCTATGTGGTCAAGACCAAGGAACGGCCCTGGTGGCCCGACAACGCCATCAGGAGATGGTGGATCAGCGTTGCTATCGTCATCTTCATCGGAGTCTGTCT 10000

  V  A  L  M  Y  F  T  Q  R  Q  A  Q  S  T  N  G  G  S  S  G  -
GGTGGCCCTGATGTACTTTACGCAGCGGCAAGCGCAGAGCACCAACGGCGGCAGCAGCGGCTAGACAAGTTTGTGGCGGCTACAGCTCCAAGCGCCGTAG 10100
                                                                                -   L  E  L  R  R  L

CCGGCCCGCCTGCCGATCGCGACGTCGTGGAGCATCGAACAGAGACTCACGCGTACGAGACCTCGAGGTACGCCACGCGGTGCCTAACGCGGTATACCAC 10200
 R  G  A  Q  R  D  R  R  R  P  A  D  F  L  S  E  R  T  R  S  R  S  T  R  W  A  T  G  L  A  T  Y  W

ACCCGTACGGTCTGCAGTGCGGCGTACAACGTGTGGAAAAGGCGTCGTGTCGCAGAGTCCGCCACGTCCCTGTCTTGTCGCTCCCCAATCGGCTCCCGCA 10300
V  R  V  T  Q  L  A  A  Y  L  T  H  F  L  R  R  T  A  S  D  A  V  D  R  D  Q  R  E  G  I  P  E  R  V

CACCCCCCGCGGCACCCAGAGGGCGGGTGAGCCAAGTATTCTTAAGGCCGTTCTCTGTTGCATAGTCCATAAATTGTTGATTCCGGAGCTCGTTGGCGCG 10400
  G  G  A  A  G  L  P  R  T  L  W  T  N  K  L  G  N  E  T  A  Y  D  M  F  Q  Q  N  R  L  E  N  A  R

GAAATAGCCGGATAAGGGGAGCAACAACCGTCGGCGAAAGCCGTCCTGCTCATTCAGTCCGGGTTTTGCGTCCAGTCGGACGTGTGACCGTTGGGCAACG 10500
 F  Y  G  S  L  P  L  L  L  R  R  R  F  G  D  Q  E  N  L  G  P  K  A  D  L  R  V  H  S  R  Q  A  V

GAACGGCGTTTCACTGCCAAAATCGTATCGCGTAGTGTACGAGACGTCGACAGTGTAGAATGCGACTCGCGGCGTAGCTCGCCGTCGCTATGCGGCTCGT 10600
S  R  R  K  V  A  L  I  T  D  R  L  T  R  S  T  S  L  T  S  H  S  E  R  R  L  E  G  D  S  H  P  E  D

CGCCGTGTGGCGCGGCCTGGCCGGCTGTCTGCGTCCAGATCTGTTGGCTTTTTGGTTTCTCTGGCTGCTGCTGCGTGTGTGCTTTGGCAGACGCGGTGGC 10700
  G  H  P  A  A  Q  G  A  T  Q  T  W  I  Q  Q  S  K  P  K  E  P  Q  Q  Q  T  H  A  K  A  S  A  T  A

AGTGTGTGGTCTGCGGTAAGTGAGGATGTCGCCGAGCAGGCGCACTTGCGGCGCGTGGGCGGCACGCGTGTTATTGTAGGTTCGTTGCCAGATGGCAAAT 10800
 T  H  P  R  R  Y  T  L  I  D  G  L  L  R  V  Q  P  A  H  A  A  R  T  N  N  Y  T  R  Q  W  I  A  F

GCTGTCGACAGCAGACGTGGGCGGTCGGTGTATTTTTGTGGGTTGCGGTGAAAGTCGGCAGTCGGTGTTTTGAGAGTCATCTTAACCATCTGTGTTGCTT 10900
A  T  S  L  L  R  P  R  D  T  Y  K  Q  P  N  R  H  F  D  A  T  P  T  K  L  T  M  K  V  M  Q  T  A  K

TGAGCAGCGTCCAGAACAGCGACGCGACTTTGGGGATGGCCTCGTGCTCACCTCCGCGGAGAGCGCCGCCGGACCTGCTCGTCAGCAGCGAGCTACGCAG 11000
  L  L  T  W  F  L  S  A  V  K  P  I  A  E  H  E  G  G  R  L  A  G  G  S  R  S  T  L  L  S  S  R  L

ACGGAATATCTGGAGGAGAGTTACGTGTGTCACAGGAGAGCGCGGGTCACCGGCGGTAACGACGGCGGTGTCGTCGACACGTGTGCGGCCTGTTGTGCTC 11100
 R  F  I  Q  L  L  T  V  H  T  V  P  S  R  P  D  G  A  T  V  V  A  T  D  D  V  R  T  R  G  T  T  S

TGCGGAAAAGTGCCGGTCTTGGAGATCGTGGACGAAAAAGAGAACGCAGCAGCTACCGCTGGCGGCGGCGGCGTTAATGCAGCCGTTGATGTTCGACGTT 11200
Q  P  F  T  G  T  K  S  I  T  S  S  F  S  F  A  A  A  V  A  P  P  P  P  T  L  A  A  T  S  T  R  R  Q

GTGAGTACTCGGAAACAGCGGTGAGGCAGAAGGTCGATCCTCCAGGGAACGACAGTCGATGCGTGGTAGCTGCAGCAGGTGAGGTTGGGGCGGACAACGT 11300
  S  Y  E  S  V  A  T  L  C  F  T  S  G  G  P  F  S  L  R  H  T  T  A  A  A  P  S  T  P  A  S  L  T

GTTGCGGATCGTGGCGAGAACGTCGTCCTCCCCTTCTTCACCGCCCCACCCACCCTCGGTTTGTGTTTCTTTTTTCTTGTGTTCTGTAGATAGTTCCATG 11400
 N  R  I  T  A  L  V  D  D  E  G  E  E  G  G  W  G  G  E  T  Q  T  E  K  K  K  H  E  T  S  L  E  M

GACAGCGACGGCAAGTCCATAATCACCGGTGTGCAAGTGGTGGAACACGACGAAGATATCATAGCGCCGCAGAGTTTGTGGTGCACGGCGTTCAAGGAAG 11500
S  L  S  P  L  D  M  I  V  P  T  C  T  T  S  C  S  S  S  I  M  A  G  C  L  K  H  H  V  A  N  L  S  A

CCCTCTGGGATGTGGCTCTGTTGGAAGTGCCGCGTTGGGCGTGGCAGGGCTGGAAGAGGTGGCGCAACAGCGAGTCCGGGCGTCGGTGGAGTGCTGGGTC 11600
  R  Q  S  T  A  R  N  S  T  G  R  Q  A  H  C  P  Q  F  L  H  R  L  L  S  D  P  R  R  H  L  A  P  D

CGCGTCGGCCTCCAGCTTGTCTGACTTGGCGGGCGAGGCCGTTGGAGAATTGGTGGGATCGGTCGTCGCGTACGTGATCCTTGAACGTCTGTGGTTGGCA 11700
 A  D  A  E  L  K  D  S  K  A  P  S  A  T  P  S  N  T  P  D  T  T  A  Y  T  I  R  S  R  R  H  N  A

GCCAGAGGCTGGGTGTGCGAAACAGGTGTGGAAGCCGAGGAGGCCATGGTGCGGCGGCGACAGCGCATGCTGTGGCGTATGTTCTCTCGTGGAGGCGACG 11800
A  L  P  Q  T  H  S  V  P  T  S  A  S  S  A  M  T  R  R  R  C  R  M  S  H  R  I  N  E  R  P  P  S  P

GCGAATGCAGCAGACGGTGTTCGATGGAGATGGCGCGCGAGGAAGAAAGCGCCGTGTTGTGAGCAGACGACGTTGGATGCGGGACGTCGGAGCAGATGGG 11900
  S  H  L  L  R  H  E  I  S  I  A  R  S  S  S  L  A  T  N  H  A  S  S  T  P  H  P  V  D  S  C  I  P

CCATGTGTGGTGGCAGATGGCGGTGTCCACTTGTGCCTGTCGCGGTAGTGCACAGACGAAGCAACATGTCGTTGTGAAGAGATAGAGTGAGAGCATAGCT 12000
  M  H  P  P  L  H  R  H  G  S  T  G  T  A  T  T  C  L  R  L  L  M  UL150

GTATGCAGCGTTGTGTGTGGAAGCGGGGGGAATAAGACGTTAATAAAGAATAGCGGCGGTTCTGAGAGGGCGACCGCTGAAGCGAGTTGCGTGTGCGTGC 12100

GGTTTGTGGTTCGAAGCGCAAAAGGCCCCCGGTCCCGCACATCCTCCGTCCCCGCAGGAGGCCTCGTCGCGGCCGCAAACTCTCCCCCGTCCCCGCACAC 12200
```

Figure 5.1 The nucleotide sequence of the right end of U$_L$ in AD169*var*UC
(Continued overleaf)

```
CCCCGTCCGCGCCGCAAACTGTCCCCGTCCCCAACGTAACCTCCCCGACGCGGCGCGAACAGCCCCGGCCCCAGCGCAACCCCCGTCCCCGGCCCCAACA 12300

CCGTCCCGCACACCCCCCGTCTCCGCAACACCCCGGCAGCGCCGGCGGCCAGAACGCTCGAAAACCCCCGAGAAGCGCAGCGCCGAAACGACACAGGCAA 12400

GGACCGTGGAACGCACCGGCAGCGCGCCGAAACACCGTCCCGAAGCCCGGTGCCGACAACAAATACCGTGGGACGACACGCACCGGCAGTGCGCAGGCAG 12500

CGGCGGACACAACACGCTTACGGCCCTCAACACTCCCTCGAGGACCCACCACGCGGCCCGGAATGGACCACGCGGCCTCAGCCGGCGGTGTTTTGGGTGT 12600

GTCGGGGCGCGGCCGGGTGGGTGTGTGCCGGGTGTGTCGCGGGCGTGTGTTGGGTGTGTCGGGGGTGTGTTGGCAGGGTGTGTCAGGGTGTGTCGCGGGC 12700

GTGTGCCGGGTGTGTCGTGCCGGGTGTGTCGCGGGCGTGTGGCGGGTGTGCCGGCGGGGTGTGGTGGCGGGGTGTGTCGGCGGTGTGCGCGGCCTCGGGG 12800

TGTGCGGCTTCGCAGGAACGAGTGTGTGGCCTCGCGGCCGTTATTTCCCCCGCGGTCCCCAGGGCCGTCGTCCCTCGCCCCCGGGCGTTGCTTTTCGTGT 12900

GTCCCCAGGGACCCATGCTGCCGTCCCCCGGGAACTTCCTCTTTTCCCCGGGGAATCACACAGACACAGACACGCGTCTTCTTTTCGCCGTGCGCGCCGC 13000

ACGTCGCTTTTATTCGCCGTCGCCGTCCTCCGCACCACACGCAACTAGTCGCCGTCCACACACGCAACTCCAAGTTTCACCCCCCCGCTAAAAACACCCC 13100

CCCGCCCCTCGAGGACCCACCACGCGGCCCGGAATGGATGTCGGGCGTCCACCTAGATGGGTGCGCGCCCGGGAGGCGGCTGTGCGCTCCAGTGGTACGC 13200

GCCTGCCGCGCGTCTTCCTTCGGGTAGCTGCCTTTCCCAGTCCACGGCCTTCCAGACTGCGTGGCGCCAAGGCGGCGCCAGCACGCGCCGTGCACGTCGC 13300

TGCCTATAAAAGCCAGCTGCGTGTCGCCCGCGGCACACGGGCGACGAAGGCGTCCGCGTGTCTAAACCGCGTGCTCGCTGACGCGGGTTTGCTTCCTATA 13400

                             M  A  Q  R  N  G  M  S  P  R  P  P  P  L  G  R  G  R  G  A  G  G  P  S
TAGTGGACGTCGGAGGTGTCCGGCGCCCATGGCCCAGCGCAACGGCATGTCGCCGCGCCCCCGCCCCTTGGTCGCGGCCGCGGGGCCGGAGGGCCTTCG 13500
                            IRS1/TRS1
 G  V  G  S  S  P  P  S  S  C  V  P  M  G  A  P  S  T  A  G  T  G  A  S  A  A  A  T  T  T  P  G  H
GGGGTTGGTTCCTCTCCTCCTTCTTCTTGTGTGCCGATGGGAGCGCCGTCCACAGCGGGCACTGGTGCGAGTGCTGCGGCTACGACGACGCCGGGCCACG 13600
G  V  H  R  V  E  P  R  G  P  P  G  A  P  P  S  S  G  N  N  S  N  F  W  H  G  P  E  R  L  L  L  S  Q
GCGTCCACCGGGTAGAACCCCGCGGGCCGCCGGGCGCCCCTCCGAGTAGCGGCAACAATAGCAACTTTTGGCACGGCCCGGAGCGCCTGTTGCTGTCTCA 13700

  I  P  V  E  R  Q  A  L  T  E  L  E  Y  Q  A  M  G  A  V  W  R  A  A  F  L  A  N  S  T  G  R  A  M
GATTCCGGTGGAGCGCCAGGCGCTGACGGAGCTGGAATACCAGGCCATGGGCGCCGTGTGGCGCGCGGCGTTTTTGGCCAACAGCACGGGCCGCGCCATG 13800

 R  K  W  S  Q  R  D  A  G  T  L  L  P  L  G  R  P  Y  G  F  Y  A  R  V  T  P  R  S  Q  M  N  G  V
CGCAAGTGGTCGCAGCGCGACGCGGGCACGCTGCTGCCGCTCGGACGGCCGTACGGATTCTACGCGCGGGTGACGCCGCGCAGCCAGATGAACGGCGTGG 13900

G  A  T  D  L  R  Q  L  S  P  R  D  A  W  I  V  L  V  A  T  V  V  H  E  V  D  P  A  A  D  P  T  L  G
GCGCGACGGACCTGCGTCAACTGTCGCCGCGGGACGCGTGGATCGTACTGGTGGCTACCGTGGTGCACGAGGTGGACCCCGCAGCCGACCCGACGTTGGG 14000

  D  K  A  G  H  P  E  G  L  C  A  Q
CGACAAGGCCGGCCATCCCGAGGGTCTGTGCGCGCAG                                                                14037
```

Figure 5.1 The nucleotide sequence of the right end of U_L in AD169*var*UC

The nucleotide sequence of the right end of U_L in AD169*var*UC and its encoded amino acid sequences, with the co-ordinates on the right. Putative start codons are highlighted in blue and putative stop codons are highlighted in green. The right genome end of U_L in AD169*var*UC is highlighted in red and the left genome end is highlighted in yellow. The location of the 3.2 kbp deletion is highlighted in pink. The alternative C terminus of UL140 is highlighted in grey.

Those strains that showed the highest level of sequence identity with AD169*var*UC are displayed in Table 5.3. No previously sequences strain is identical to AD169*var*UC. As in other strains of HCMV, the rightmost gene of U_L in AD169*var*UC is UL150, after which the sequence continues into R_L and subsequently R_S (Dolan *et al.*, 2004; Murphy *et al.*, 2003). This confirms that AD169*var*UC does not contain the inverted duplicated region (containing RL1-RL12 and part of RL13) that is present in AD169*var*UK and AD169*var*ATCC.

Table 5.3: Comparison of AD169*var*UC ORFs in U_L/*b'* with homologous sequences in other HCMV strains

ORF	*Length of ORF (bp)* [a]	*Sequence identity* [b] *(%)*	*HCMV strain* [c]
UL139	441	99	W
Partial UL140	548	98.7	U3
		98	Towne
		97	W
Partial UL144	375	100	Towne
UL145	393	100	Towne
UL146	357	100	FS
UL147	480	100	CH25
		95	Towne
UL147A	228	100	CH25
		91	Towne

[a] Includes the stop codon.

[b] Sequence identity between AD169*var*UC and other HCMV strains.

[c] HCMV strains with the highest level of identity to the homologous ORF in AD169*var*UC.

5.4 Sequence of the inverted repeat regions and left end of U_L in AD169*var*UC

In addition to sequencing U_L/*b'*, the inverted repeat regions (TR_L/TR_S and IR_L/IR_S) were also sequenced. TR_L and the part of TR_S sequenced proved to be identical to the corresponding regions of IR_L/IR_S in AD169varUC. Figure 5.2 shows a nucleotide alignment of R_S/R_L and the flanking sequence at the left end of U_L in the three variants. The comparisons revealed that 24 nucleotides differ, with AD169*var*UC differing from AD169*var*UK at 8 positions, AD169*var*UC differing from AD169*var*ATCC at 23 positions, and AD169*var*UK differing from AD169*var*ATCC at 17 positions.

Also, there is a region of extensive difference in the *a* sequence between the three AD169 variants. This is largely due to an additional 492 bp in AD169*var*ATCC, which is not present in AD169*var*UK or AD169*var*UC. It is likely to be due to duplication of a repeat element. In addition, AD169*var*UC contains an additional 18 bp not found in

AD169*var*UK and AD169*var*UC differs from AD169*var*UK in two of four nucleotides immediately downstream of this 18 bp insertion.

5.5 Discussion

Sequencing of a number of genes confirmed that AD169*var*UC is indeed a variant of AD169. In an initial investigation, the sequences of RL13, UL11, UL73, UL131A and the partial sequence of UL148 in AD169*var*UC were shown to be 100% identical to their equivalents in AD169*var*UK and AD169*var*ATCC (Table 5.1). In further studies, UL121, the main exon of UL122, a portion of the main exon of UL123, UL128, UL130, UL131A and UL132 were shown to be 100% identical to AD169*var*UK and AD169*var*ATCC (Table 5.1). Sequencing of RL5A in AD169*var*UC revealed that it has a single nucleotide mismatch with AD169*var*UK and AD169*var*ATCC. As RL5A is mutated in all three strains and probably non-functional (Davison *et al.,* 2003) this difference is unlikely to have any affect.

The sequence of U_L/*b'* in AD169*var*UC revealed the presence of 15 genes (UL148, UL147A, UL147, UL146, UL145, UL139, UL138, UL136, UL135, UL133, UL148A, UL148B, UL148C, UL148D and UL150) that are absent from AD169*var*UK and AD169*var*ATCC (Figure 5.1). However, AD169*var*UC contains only part of UL144 and UL140. It has undergone a 3.2 kbp deletion that results in deletion of the complete UL141 and UL142 ORFs, the first 148 bp of UL144 and last 27 bp of UL140. The residual portion of UL144 is unikely to be functional. It is not known whether the frameshifted variant of UL140 is functional.

A comparison of the partial sequence of UL140, as well as the full sequences of genes on either side of the deletion, with sequences from other HCMV strains revealed that some genes are identical to those in one strain, whereas others are identical to those in other strains (Table 5.3). This result is not surprising, given the general lack of linkage between hypervariable genes across the HCMV genome (Rasmussen *et al.,* 2003) and the length of the deletion (3.2 kbp). Even overlooking the deletion, the sequence of U_L/*b'* in AD169varUC differs from other sequenced strains.

```
varUC     CTGCGCGCACAGACCCTCGGGATGGCCGGCCTTGTCGCCCAACGTCGGGTCGGCTGCGGGGTCCACCTCGTGCACCACGGTAGCCACCAGTACGATCCAC
varUK     CTGCGCGCACAGACCCTCGGGATGGCCGGCCTTGTCGCCCACCGTCGGGTCGGCTGCGGGGTCCACCTCGTGCACCACGGTAGCCACCAGTACGATCCAC
varATCC   CTGCGCGCACAGACCCTCGGGATGGCCGGCCTTGTCGCCCACCGTCGGGTCGGCTGCGGGGTCCACCTCGTGCACCACGGTAGCCACCGGTACGATCCAC
```

R_S

```
varUC     GCGTCCCGCGGCGACAGTTGACGCAGGTCCGTCGCGCCCACGCCGTTCATCTGGCTGCGCGGCGTCACCCGCGCGTAGAATCCGTACGGCCGTCCGAGCG
varUK     GCGTCCCGCGGCGACAGTTGACGCAGGTCCGTCGCGCCCACGCCGTTCATCTGGCTGCGCGGCGTCACCCGCGCGTAGAATCCGTACGGCCGTCCGAGCG
varATCC   GCGTCCCGCGGCGACAGTTGACGCAGGTCCGTCGCGCCCACGCCGTTCATCTGGCTGCGCGGCGTCACCCGCGCGTAGAATCCGTACGGCCGTCCGAGCG

varUC     GCAGCAGCGTGCCCGCGTCGCGCTGCGACCACTTGCGCATGGCGCGGCCCGTGCTGTTGGCCAAAAACGCCGCGCGCCACACGGCGCCCATGGCCTGGTA
varUK     GCAGCAGCGTGCCCGCGTCGCGCTGCGACCACTTGCGCATGGCGCGGCCCGTGCTGTTGGCCAAAAACGCCGCGCGCCACACGGCGCCCATGGCCTGGTA
varATCC   GCAGCAGCGTGCCCGCGTCGCGCTGCGACCACTTGCGCATGGCGCGGCCCGTGCTGTTGGCCAAAAACGCCGCGCGCCACACGGCGCCCATGGCCTGGTA

varUC     TTCCAGCTCCGTCAGCGCCTGGCGCTCCACCGGAATCTGAGACAGCAACAGGCGCTCCGGGCCGTGCCAAAAGTTGCTATTGTTGCCGCTACTCGGAGGG
varUK     TTCCAGCTCCGTCAGCGCCTGGCGCTCCACCGGAATCTGAGACAGCAACAGGCGCTCCGGGCCGTGCCAAAAGTTGCTATTGTTGCCGCTACTCGGAGGG
varATCC   TTCCAGCTCCGTCAGCGCCTGGCGCTCCACCGGAATCTGAGACAGCAACAGGCGCTCCGGGCCGTGCCAAAAGTTGCCATTGTTGCCGCTACTCGGAGGG

varUC     GCGCCCGGCGGCCCGCGGGGTTCTACCCGGTGGACGCCGTGGCCCGGCGTCGTCGTAGCCGCAGCACTCGCACCAGTGCCCGCTGTGGACGGCGCTCCCA
varUK     GCGCCCGGCGGCCCGCGGGGTTCTACCCGGTGGACGCCGTGGCCCGGCGTCGTCGTAGCCGCAGCACTCGCACCAGTGCCCGCTGTTGACGGCGCTCCCA
varATCC   GCGCCCGGCGGCCCGCGGGGTTCTACCCGGTGGACGCCGTGGCCCGGCGTCGTCGTAGCCGCAGCACTCGCACCAGTGCCCGCTGTGGACGGCGCTCCCA

varUC     TCGGCACACAAGAAGAAGGAGGAGAGGAACCAACCCCCGAAGGCCCTCCGGCCCCGCGGCCGCGACCAAGGGGCGGGGGGCGCGGCGACATGCCGTTGCG
varUK     TCGGCACACAAGAAGAAGGAGGAGAGGAACCAACCCCCGAAGGCCCTCCGGCCCCGCGGCCGCGACCAAGGGGCGGGGGGCGCGGCGACATGCCGTTGCG
varATCC   TCGGCACACAAGAAGAAGGAGGAGAGGAACCAACCCCCGAAGGCCCTCCGGCCCCGCGGCCGCGACCGAGGGGCGGGGGGCGCGGCGACATGCCGTTGCG

varUC     CTGGGCCATGGGCGCCGGACACCTCCGACGTCCACTATATAGGAAGCAAACCCGCGTCAGCGAGCACGCGGTTTAGACACGCGGACGCCTTCGTCGCCCG
varUK     CTGGGCCATGGGCGCCGGACACCTCCGACGTCCACTATATAGGAAGCAAACCCGCGTCAGCGAGCACGCGGTTTAGACACGCGGACGCCTTCGTCGCCCG
varATCC   CTGGGCCATGGGCGCCGGACACCTCCGACGTCCACTATATAGGAAGCAGACCCGCGTCAGCGAGCACGCGGTTTAGACACGCGGACGCCTTCGTCGCCCG

varUC     TGTGCCGCGGGCGACACGCAGCTGGCTTTTATAGGCAGCGACGTGCACGGCGCGTGCTGGCGCCGCCTTGGCGCCACGCAGTCTGGAAGGCCGTGGACTG
varUK     TGTGCCGCGGGCGACACGCAGCTGGCTTTTATAGGCAGCGACGTGCACGGCGCGTGCTGGCGCCGCCTTGGCGCCACGCAGTCTGGAAGGCCGTGGACTG
varATCC   TGTGCCGCGGGCGACACGCAGCTGGCTTTTATAGGCAGCGACGTGCACGGCGCGTGCTGGCGCCGCCTTGGCGCCACGCAGTCTGGAAGGCCGTGGACTG

varUC     GGAAAGGCAGCTACCCGAAGGAAGACGCGCGGCAGGCGCGTACCACTGGAGCGCACAGCCGCCTCCCGGGCGCGCACCCATCTAGGTGGACGCCCGACAT
varUK     GGAAAGGCAGCTACCCGAAGGAAGACGCGCGGCAGGCGCGTACCACTGGAGCGCACAGCCGCCTCCCGGGCGCGCACCCATCTAGGTGGACGCCCGACAT
varATCC   GGAAAGGCAGCTACCCGAAGGAAGACGCGCGGCAGGCGCGTACCACTGGAGCGCACAGCCGCCTCCCGGGCGCGCACCCATCTAGGTGGACGCCCGACAT

varUC     CCATTCCGGGCCGCGTGGTGGGTCCTCGAGGGGCGGGGGGGTGTTTTTAGCGGGGGGGTGAAACTTGGAGTTGCGTGTGTGGACGGCGACTAGTTGCGTG
varUK     CCATTCCGGGCCGCGTGGTGGGTCCTCGAGGGGCGGGGGGGTGTTTTTAGCGGGGGGGTGAAACTTGGAGTTGCGTGTGTGGACGGCGACTAGTTGCGTG
varATCC   CCATTCCGGGCCGCGTGGTGGGTCCTCGAGGGGCGGGGGGGTGTTTTTAGCGGGGGGGTGAAACTTGGAGTTGCGTGTGTGGACGGCGACTAGTTGCGTG
```

Start of the *a* sequence

```
varUC     TGGTGCGGAGGACGGCGACGGCGAATAAAAGCGACGTGCGGCGCGCACGGCGAAAAGAAGACGCGTGTCTGTGTCTGTGTGATTCCCCGGGGAAAAGAGG
varUK     TGGTGCGGAGGACGGCGACGGCGAATAAAAGCGACGTGCGGCGCGCACGGCGAAAAGAAGACGCGTGTCTGTGTCTGTGTGATTCCCCGGGGAAAAGAGG
varATCC   TGGTGCGGAGGACGGCGACGGCGAATAAAAGCGACGTGCGGCGCGCACGGCGAAAAGAAGACGCGTGTGTGTGTCTGTGTGATTCCCCGGGGAAAAGAGG

varUC     AAGTTCCCGGGGGACGGCAGCATGGGTCCCTGGGGACACACGAAAAGCAACGCCCGGGGGCGAGGGACGACGGCCCTGGGGACCGCGGGGGAAATAACGG
varUK     AAGTTCCCGGGGGACGGCAGCATGGGTCCCTGGGGACACACGAAAAGCAACGCCCGGGGGCGAGGGACGACGGCCCTGGGGACCGCGGGGGAAATAACGG
varATCC   AAGTTCCCGGGGGACGGCAGCATGGGTCCCTGGGGACACACGAAAAGCAACGCCCGGGGGCGAGGGACGACGGCCCCGGGGACCGCGGGGGAAATAACGG

varUC     CCGCGAGGCCACACACTCGTTCCTGCGAAGCCGCACACCCCGAGGCCGCGCACACCGCCGACACACCCCGCCACCACACCCCGCCGGCACACCCGCCACA
varUK     CCGCGAGGCCACACACTCGTTCCTGCGAAGCCGCACACCCCGAGGCCGCGCACACCGCCGACACACCCCGCCACCACACCCCGCCGGCACACCCGCCACA
varATCC   CCGCGAGGCCACACACTCGTTCCCGCGAAGCCGCACACCCCGAGGCCGCGCACACCGCCGACACACCCCGCCACCACACCCCGCCGGCACACCCGCCACA

varUC     CGCCCGCGACACACCCGGCACGACACACCCGGCACACGCCCGCGACACACCCTGACACACCCTGCCAACACACCCCCGACACACCCAACACACGCCCGCG
varUK     CGCCCGCGACACACCCGGCACGACACACCCGGCACACGCCCGCGACACACCCTGACACACCCTGCCAACACACCCCCGACACACCCAACACACGCCCGCG
varATCC   CGCCCGCGACACACCCGGCACGACACACCCGGCACACGCCCGCGACACACCCTGACACACCCTGCCAACACACCCCCGACACACCCAACACACGCCCGCG

varUC     ACACACCCGGCACACACCCACCCGGCCGCGCCCCGACACACCCAAAACACCGCCGG............................................
varUK     ACACACCCGGCACACACCCACCCGGCCGCGCCCCGACACACCCAAAACACCGCCGG............................................
varATCC   ACACACCCGGCACACACCCACCCGGCCGCGCCCCGACACACCCAAAACACCGCCGGTCCATTCCGGGCCGCCCATTCCGGGCCGCGTGGTGGGTCCATTC

varUC     ....................................................................................................
varUK     ....................................................................................................
varATCC   CGGGCCGCGTGGTGGGTCCATTCCGGGCCGCGTGGTGGGTCCTCGAGGGAGTGTTGAGGGCCGTAAGCGTGTTGTGTCCGACGCTGCCTGCGCACTGCCG

varUC     ....................................................................................................
varUK     ....................................................................................................
varATCC   GTGCGTGTCGTCCCACGGTATTTGTTGTCGGCACCGGGCTTCGGGACGGTGTTTCGGCGCGCTGCCGGTGCGTTCCACGGTCCTTGCCTGTGTCGTTTCC

varUC     ....................................................................................................
varUK     ....................................................................................................
varATCC   GGCCGCGCCCCGACACACCCAAAACACCGACGTGCGGGGCCGCGTGGTGGGTCCTCGAGGGAGTGTTGAGGGCCGTAAGCGTGCTGTGTCCGACGCTGCC
```

Figure 5.2 A nucleotide alignment of R_L/R_S in AD169*var*UK, AD169*var*ATCC and AD169*varUC* (continued overleaf)

```
varUC       ..........................CTGAGGCCGCGTGGTCCATTCCGGGCCGCGTGGTGGGTCCTCGAGGGAGTGTTGAGGGCCGTAAGCGTGTTGTG
varUK       ............................................TGCGGGGCCGCGTGGTGGGTCCTCGAGGGAGTGTTGAGGGCCGTAAGCGTGTTGTG
varATCC     TGTGTCGTTTCCGGCCGCG..............TGGTGGGTCCATTCCGGGCCGCGTGGTGGGTCCTCGAGGGAGTGTTGAGGGCCGTAAGCGTGTTGTG
varATCCalt  TGTGTCGTTTCCGGCCGCGCCCCGACACACCCAAAACACCGACGTGCGGGGCCGCGTGGTGGGTCCTCGAGGGAGTGTTGAGGGCCGTAAGCGTGTTGTG
```

end of the *a* sequence

```
varUC       TCCGCCGCTGCCTGCGCACTGCCGGTGCGTGTCGTCCCACGGTATTTGTTGTCGGCACCGGGCTTCGGGACGGTGTTTCGGCGCGCTGCCGGTGCGTTCC
varUK       TCCGACGCTGCCTGCGCACTGCCGGTGCGTGTCGTCCCACGGTATTTGTTGTCGGCACCGGGCTTCGGGACGGTGTTTCGGCGCGCTGCCGGTGCGTTCC
varATCC     TCCGACGCTGCCTGCGCACTGCCGGTGCGTGTCGTCCCACGGTATTTGTTGTCGGCACCGGGCTTCGGGACGGTGTTTCGGCGCGCTGCCGGTGCGTTCC

varUC       ACGGTCCTTGCCTGTGTCGTTTCGGCGCTGCGCTTCTCGGGGGTTTTCGAGCGTTCTGGCCGCCGGCGCTGCCGGGGTGTTGCGGAGACGGGGGGTGTGC
varUK       ACGGTCCTTGCCTGTGTCGTTTCGGCGCTGCGCTTGTCGGGGGTTTTCGAGCGTTCTGGCCGCCGGCGATGCCGGGGTGTTGCGGAGACGGGGGGTGTGC
varATCC     ACGGTCCTTGCCTGTGTCGTTTCGGCGCTGCGCTTGTCGGGGGTTTTCGAGCGTTCTGGCCGCCGGCGATGCCGGGGTGTTGCGGAGACGGGGGGTGTGC

varUC       GGGACGGTGTTGGGGCCGGGGACGGGGGTTGCGCTGGGGCCGGGGCTGTTCGCGCCGCGTCGGGGAGGTTACGTTGGGGACGGGGACAGTTTGCGGCGCG
varUK       GGGACGGTGTTGGGGCCGGGGACGGGGGTTGCGCTGGGGCCGGGGCTGTTCGCGCCGCGTAGGGGAGGTTACGTTGGGGACGGGGACAGTTTGCGGCGCG
varATCC     GGGACGGTGTCGGGGCCGGGGACGGGGGTTGCGCTGGGTCCGGGGCTGTTCGCGCCGCGTAGGGGAGGTTACGTTGGGTACGGGGACAGTTTGCGGCGCG

varUC       GACCAGGGAACCCACCTCACCTATTTAACCTCCACCCACTCCAACACACACATGCCGCACAATCATGCCAGCCACAGACACAAACAGCACCCACACCACG
varUK       GACCAGGGAACCCACCTCACCTATTTAACCTCCACCCACTACAACACACACATGCCGCACAATCATGCCAGCCACAGACACAAACAGCACCCACACCACG
varATCC     GACCAGGGAACCCACCTCACCTATTTAACCTCCACCCACTACAACACACACATGCCGCACAATCATGCCAGCCACAGACACAAACAGCACCCACACCACG
```

Start of U$_L$

```
varUC       CCGCTTCACCCAGACGCCCAACACACGTTACCCTTACACCACAGCAACACACAACCGCATGTCCAAACCTCGGACAAACACGCCGACGAAGAACACCGCA
varUK       CCGCTTCACCCAGACGCCCAACACACGTTACCCTTACACCACAGCAACACACAACCGCATGTCCAAACCTCGGACAAACACGCCGACGAAGAACACCGCA
varATCC     CCGCTTCACCCAGACGCCCAACACACGTCACCCTTACACCACAGCAACAGACAACCGCATGTCCGAACCTCGGACAAACACGCCGACGAAGAACACCGCA

varUC       CACAGATGGAGCTCGACGCCGCAGACTACGCTGCTTGCGCACAGGCCCGCCAACACCTCTACGATCAAACACAACCCCTACTACTCGCATACCCCAACAC
varUK       CACAGATGGAGCTCGACGCCGCAGACTACGCTGCTTGCGCACAGGCCCGCCAACACCTCTACGATCAAACACAACCCCTACTACTCGCATACCCCAACAC
varATCC     CACAGATGGAGCTCGACGCCGCAGACTACGCTGCTCGCGCACAGGCCCGCCAACACCTCTACGATCAAACACAACCCCTACTACTCGCATACCCCAACAC

varUC       CAACCCACAGGACAGCGCTCATTTTCCCACAGAGAATCACCATCAACTCACGCATCCACTTCACAACATTGGCGAGGGCGCAGCACTCGGCTACCCCGTC
varUK       CAACCCACAGGACAGCGCTCATTTTCCCACAGAGAATCAACATCAACTCACGCATCCACTTCACAACATTGGCGAGGGCGCAGCACTCGGCTACCCCGTC
varATCC     CAACCCACAGGACAGCGCTCATTTTCCCACAGAGAATCAACATCAACTCACGCATCCACTTCACAACATTGGCGAGGGCGCAGCACTCGGCTACCCCGTC

varUC       CCCCGCGCGGAAATCCGCCGCGGCGGTGGCGACTGGGCCGACAGCGCAAGCGACTTTGACGCCGACTGCTGGTGCATGTGGGGACGCTTCGGAACCATGG
varUK       CCCCGCGCGGAAATCCGCCGCGGCGGTGGCGACTGGGCCGACAGCGCAAGCGACTTTGACGCCGACTGCTGGTGCATGTGGGGACGCTTCGGAACCATGG
varATCC     CCCCGCGCGGAAATCCGCCGCGGCGGTGGCGACTGGGCCGACAGCGCAAGCGACTTTGACGCCGACTGCTGGCGCATGTGGGGACGCTTCGGAACCATGG

varUC       GCCGCCAACCTGTCGTCACCTTACTGTTGGCGCGCCAACGCGACGGCCTCGCTGACTGGAACGTCGTACGCTGCCGCGGCACAGGCTTTCGCGCACACGA
varUK       GCCGCCAACCTGTCGTCACCTTACTGTTGGCGCGCCAACGCGACGGCCTCGCTGACTGGAACGTCGTACGCTGCCGCGGCACAGGCTTTCGCGCACACGA
varATCC     GCCGCCAACCTGTCGTCACCTTACTGTTGGCGCGCCAACGCGACGGCCTCGCTGACTGGAACGTCGTACGCTGCCGCGGCACAGGCTTTCGCGCACACGA

varUC       TTCCGAGGACGGCGTCTCTGTCTGGCGTCAGCACCTGGTTTTTTTACTCGGAGGCCACGGCCGCCGTGTACAGTTAGAACGTCCATCCGCGGGAGAAGCC
varUK       TTCCGAGGACGGCGTCTCTGTCTGGCGTCAGCACCTGGTTTTTTTACTCGGAGGCCACGGCCGCCGTGTACAGTTAGAACGTCCATCCGCGGGAGAAGCC
varATCC     TTCCGAGGACGGCGTCTCTGTCTGGCGTCAGCACCTGGTTTTTTTACTCGGAGGCCACGGCCGCCGTGTACAGTTAGAACGTCCATCCGCGGGAGAAGCC

varUC       CAAGCTCGAGGCCTCTTGCCACGCATCCGGATCACCCCCATCTCCACATCTCCACGTCGGAAACCGCCGCACCCCGCCACATCCACCGCATCGCACCACC
varUK       CAAGCTCGAGGCCTCTTGCCACGCATCCGGATCACCCCCATCTCCACATCTCCACGTCGGAAACCGCCGCACCCCGCCACATCCACCGCATCGCACCACC
varATCC     CAAGCTCGAGGCCTCTTGCCACGCATCCGGATCACCCCCATGTCCACATCTCCACGTCGGAAACCGCCGCACCCCGCCACATCCACCGCATCGCACCACC

varUC       CACATGCTTCGCCTCGGTCAGATCACACGCTTTTTCCTGTCCCATCTACACCCTCAGCCACGGTTCACAATCCCCGAAACT
varUK       CACATGCTTCGCCTCGGTCAGATCACACGCTTTTTCCTGTCCCATCTACACCCTCAGCCACGGTTCACAATCCCCGAAACT
varATCC     CACATGCTTCGCCTCGGTGAGATCACACGCTTTTTCCTGTCCCATCTACACCCTCAGCCACGGTTCACAATCCCCGAAACT
```

Figure 5.2 A nucleotide alignment of R$_L$/R$_S$ in AD169*var*UK, AD169*var*ATCC and AD169*varUC*

A nucleotide alignment of AD169*var*UK, AD169*var*ATCC and AD169*varUC* sequence beginning within R$_S$, through the *a* sequence into R$_L$ and the left end of U$_L$. Nucleotide differences are highlighted in green. A region of extensive difference is highlighted in blue; this is largely due to duplication of a repeat sequence (underlined). The sequences of IR$_S$/IR$_L$ and TR$_S$/TR$_L$ differ somewhat in AD169*var*ATCC; therefore the sequence of the other repeat (*var*ATCCalt) is included to highlight their differences. Nucleotide positions of the genome termini are highlighted in red with the left end first and the right end second; the region between is the *a* sequence. The start of U$_L$ is highlighted in pink, the start codon of RL1 is highlighted in grey and the start codon of IRS1/TRS1 is highlighted in yellow.

AD169varUC, AD169*var*UK and AD169*var*ATCC all stem from the same original stock (NIH76559) isolated by Rowe *et al.* (1956). Where sequence differences exist, AD169varUC is more similar to AD169varUK than AD169varATCC. One possibility is that AD169varUC was derived from a very early passage of AD169 (one of the first 14 passages before NIH 76559 was established) and that the 15 kbp deletion and expansion of R_L occurred at a later passage in this series (but prior to NIH 76559). After this, AD169*var*UK was established in the UK from NIH 76559, and AD169*var*ATCC was derived from NIH 76559 that had been passaged many times in a US laboratory. The mutations present in all three variants would have occurred during the early passages before AD169varUC was derived. An alternative possibility is that the 15 kbp deletion and expansion of R_L (and perhaps some of the other shared mutations) occurred later, during the derivation of AD169*var*UK, and that this virus, was subsequently shipped to the USA, passaged many times, and submitted to the ATCC. Both of these theories are speculative, given that it has not proved possible to recover details of the origin of AD169*var*ATCC. However, the characterisation of AD169*var*UC has revealed information about the genetic status of AD169 earlier in its history, specifically the sequence of most of the U_L/*b'* region.

6 Final discussion

HCMV is a complex virus and has the largest genome of any human virus. It usually results in an asymptomatic infection that is followed by life-long latency. HCMV infects a large proportion of the population worldwide, usually during childhood (Gandhi and Khanna, 2004). . However, decreased breastfeeding with improved hygiene in the developed world has resulted in a larger proportion of uninfected adults. HCMV represents a serious disease risk for immunocompromised individuals, such as AIDS patients and transplant recipients, as well as for the unborn child. It is the leading infectious cause of congenital disease, and reactivation of latent infection in the recipient or the donor organ or cells can result in transplant rejection (Zaia *et al.,* 2002). The majority of the genome is highly conserved (Murphy *et al.,* 2003; Davison *et al.,* 2003), but several genes, including UL146 and UL139, are highly variable (Dolan *et al.,* 2004; Qi *et al.,* 2006). Infection with one strain of HCMV does not necessarily provide protection against reinfection with another strain, and multiple infections are detected frequently (Boppana *et al.,* 2001).

Numerous studies have been published that investigated the genotypes of hypervariable loci in clinical isolates, with an emphasis on the relationship between genotype and disease outcome. In general there is no convincing association between genotype and clinical disease, although there have been reports of links between certain gB and UL144 genotypes and disease, specifically gB1 and retinitis in AIDS patients (Rasmussen *et al.,* 1997), gB2-4 and fatalities in transplant recipients (Fries *et al.,* 1994) and UL144 genotypes A, C and subtypes A/C, A/B with more serious disease (Arav-Boger *et al.,* 2002). In contrast, others found no evidence for any association between gB genotype and clinical disease (Aquino and Figuerido, 2000). The majority of these studies utilised small sample sizes and different patient types, and the results have proved contradictory. Mixed infections were identified in some studies, which further complicates any attempt to establish links between genotype and pathogenesis.

This thesis investigated circulating genotypes of the two most variable genes at the right end of U_L (UL146 and UL139), in a large collection of clinical isolates from geographically diverse locations in Africa, Europe, Asia and Australia.

Previously, 14 UL146 genotypes have been described (Dolan *et al.,* 2004) and a single study investigating UL139 genotypes described six groups (Qi *et al.,* 2006). The sequences of these genes were studied in 179 clinical samples and five commonly used laboratory strains. In addition, all UL146 and UL139 sequences available in public databases were included in the phylogenetic and diversity analyses. A total of 350 UL146 sequences were analysed, and all fell into the 14 genotypes described previously. A number of previous studies had investigated UL146 sequences in passaged and unpassaged clinical isolates (Stanton *et al.,* 2005; Hassan-Walker *et al.,* 2004; Lurain *et al.,* 2006; He *et al.,* 2006), but all used relatively small numbers of geographically related samples and all, with one exception (He *et al.,* 2006), examined immunocompromised patients. Therefore, the current study is the first to have examined UL146 sequences in a large group of geographically and clinically diverse samples. For UL139, 300 sequences were analysed, and all fell into eight genotypes, G1-G8. Five UL139 genotypes (G1-4, G6) correspond to the six groups (G1, G1b, G1c, G2a, G2b, G3) described by Qi *et al.* (2006) who analysed 26 clinical samples. Therefore, three new UL139 genotypes have been identified. The large number of sequences analysed during this study does not exclude the possibility that other UL146 or UL139 genotypes may be in circulation that have yet to be discovered. The use of alternative PCR primers or sequencing whole genomes in future could address this question.

Overall there was no significant association between the UL146 or UL139 genotype of a strain and its geographical origin, a conclusion that is in agreement with that drawn by Pignatelli *et al.* (2003) from the genotypes of UL73 in a panel of 223 isolates from around the world. In the present study, UL146 G10 and G11 appeared to be restricted to European samples. However, Chinese sequences available in Genbank fall into UL146 G10 and G11, indicating that these genotypes are found outside Europe. Chi-square analysis suggested a minor bias in the geographical distribution of genotypes. Specifically, UL146 G6 (*p*=0.006) and G7 (*p*=0.047) showed statistically significant differences in their genotypic distributions, as did UL139 G7 (*p*=0.006). However, this is likely to have been a consequence of small sample numbers from some areas. Although Yate’s correction was applied during the chi-square analysis, results obtained need to be treated with caution due to frequencies of zero in some cells. Europe, for which there were more samples than any other region, displayed the greatest diversity, containing all the UL146 and UL139 genotypes detected with the exception of UL146 G6, the only example of which was detected in an Asian sample. Overall, it appears probable that sufficient sample numbers would demonstrate that all genotypes are found in all regions. It remains unclear whether each genotype diverged in geographical isolation and has subsequently been transmitted worldwide, or whether the genotypes diverged during early human history and have since been maintained by the founder effect or due to geographical segregation following human migration out of Africa.

The genotypes of UL146 and UL139 appear to have evolved predominantly under constraint (purifying selection) rather than positive selection. This is despite the high level of nucleotide and aa sequence divergence between genotypes, particularly for UL146. This conclusion is in agreement with a study by Arav-Boger *et al.* (2005), who investigated the mode of selection in UL146 and UL147 sequences in 28 clinical isolates and four laboratory strains. UL146 and UL147 encode related proteins (CXC chemokines), and both appear to have evolved under constraint, UL146 at a faster rate than UL147. The deduced mode of evolution suggests that selection pressures favour retention of these genes and that they are now evolving slowly. The finding that UL146 sequences are stable *in vitro*, when passaged in cell culture, as well as *in vivo*, in samples taken from the same patient over time, is in accord with this, although the time scales involved in these studies was short on an evolutionary scale (Stanton *et al.*, 2005; Lurain *et al.*, 2006).

Published work on linkage disequilibrium between variable genes has produced positive evidence for genes that are near each other, such as UL6/UL7, UL4/UL7, UL1/UL4 and UL4/UL6 (Sekulin *et al.*, 2007), gH/gO (UL75/UL74, specifically gH1/gO1) (Rasmussen *et al.*, 2002) and gN/gO (UL73/UL74) (Mattick *et al.*, 2004). Rasmussen *et al.* (2003) found no evidence for linkage disequilibrium between six generally more widely distributed genes (UL55 (gB), UL74 (gO), UL75 (gH), UL115 (gL), US9 and US28), and concluded that genetic linkage is rare. In accordance with these findings, the present study did not detect linkage disequilibrium between UL146 and UL139 genotypes, even though the two genes are only 5.2 kbp apart on the genome. These findings support the notion that HCMV has undergone multiple recombination events during its evolution.

UL146 is a CXC chemokine that is thought to promote virus dissemination by attracting neutrophils to the initial site of infection. Penfold *et al.* (1999) detected the UL146 protein at L times p.i., which is in agreement with the transcript mapping results in the present study, where the UL146 3.3 kb mRNA was expressed with L kinetics. The 3'-end of UL146 was mapped downstream from UL132, indicating that UL146 is 3'-coterminally expressed with UL147, UL147A, UL148 and UL132. This agrees with the results of a study published during the course of this work (Lurain *et al.,* 2006), which investigated transcription of UL146 and adjacent genes by RT-PCR and northern blotting. However, in that study UL146 was characterised on an E-L transcript, albeit at low levels (due to faint band obtained). Therefore, differences between the findings of the present study (UL146 was characterised as a L gene) and those of Lurain *et al.* (2006) may be a consequence of quantitative differences.

The high level of divergence of UL146 at both the nucleotide and aa sequence levels suggests that there may also be divergence at the structural and functional levels. Nonetheless, homology modelling using the solved crystal structure of the functionally related chemokine IL-8 predicted that all 14 UL146 genotypes encode proteins with similar tertiary structures. This could indicate that, despite hypervariation, the UL146 genotypes are functionally similar to each other. However, other chemokines such as gro-α and 1F9s also share similar tertiary structures and yet display differing binding affinities for cellular receptors (Baggiolini *et al.,* 1997). Functional studies are required to determine whether this phenomenon applies to UL146 genotypes.

UL139 is predicted to encode a type I membrane glycoprotein (Cha *et al.,* 1996). No information has been published regarding UL139 function, although a region of similarity with CD24, a signal transducer involved in B cell activation, has been noted (Qi *et al.,* 2006). If this similarity is functionally significant, it would suggest a role for UL139 in regulation or modulation of the immune response. In the present study, UL139 from HCMV strain Merlin expressed in HFFF-2 cells was expressed with E-L kinetics. UL139 is 3'-coterminally expressed on a 2.6 kb mRNA transcript with UL140 and UL141.

As an initial characterisation of the UL139 protein, recombinant adenovirus vectors expressing FLAG-tagged UL139 variants (three genotypes) were produced. The tagged proteins were detected by immunoblot using an antibody against the tag. The UL139 protein was predicted to be highly glycosylated, and the preliminary experiment found the proteins to be much larger than those predicted from the unmodified aa sequences. The RADs generated in the present study will facilitate future experiments on localisation of the UL139 protein in the infected cell and virus, and on its interactions with cellular proteins.

Mixed infections with different HCMV strains were identified in 14% of the samples genotyped and this number increased to 29% when the results from repeat experiments were included. This suggests that the number of mixed infections may be underestimated. Mixed infections were detected in immunocompromised and immunocompetent individuals, suggesting that infection with multiple strains occurs in both asymptomatic and symptomatic infections. Indeed, mixed infection in immunocompromised individuals, such as transplant recipients, has been associated with enhanced pathogenesis and increased risk of transplant rejection (Coaquette *et al.,* 2004; Puchhammer-Stöckl *et al.,* 2006). The common occurrence of infection with multiple strains undermines any attempt to draw conclusions regarding association between genotype and clinical disease. Puchhammer-Stöckl *et al.* (2006) encountered a similar situation in an examination of gB (UL55) and gN (UL73) genotypes. The phenomenon of mixed infections highlights the complexity of HCMV and the problems facing HCMV vaccine development. Moreoever, samples cultured *in vitro* are likely to represent only a subset of strains present in the original clinical sample; it is important in such studies to analyse the clinical material.

AD169 is a commonly used laboratory strain of HCMV that has a large deletion (15 kbp) at the right end of U_L (Cha *et al.,* 1996), a duplication of sequences at the left end of U_L that replace the deleted region, and a number of other mutations (Akter *et al.,* 2003; Davison *et al.,* 2003; Yu *et al.,* 2002; Skaletskaya *et al.,* 2001). UL139 and UL146 are both located in this deleted region and are found in clinical material and low passage clinical isolates of HCMV such as Toledo and Merlin. Other immunomodulatory genes such as UL144, which encodes a TNF α-like receptor, are also located in this region. It is likely that the mutations observed in AD169 are a result of adaptation to serial passage in cell culture. A variant of AD169, AD169*var*UC, was acquired for which there was evidence that it contained some or all of the genes deleted in AD169*var*UK (N.

Lurain, personal communication). Initial sequencing of a number of genes (including UL146 and UL139) confirmed the identity of AD169*var*UC as derived from the original AD169 clinical material.

The right end of U_L in AD169*var*UC was sequenced and also found to be a mutant, as it contains a 3.2 kbp deletion. The deletion results in absence of UL141 and UL142, 3'-truncation of UL140 (although the C-terminal 8 aa are replaced with alternative 71 aa) and deletion of the first 148 bp of UL144. A similar sized deletion that affects the same genes (i.e. UL141, UL142, UL140 and UL144) has been noted in low-passage isolate VR1814 (A. Davison, personal communication). This suggests that this region is prone to deletion during cell culture. It is unclear whether AD169 first underwent this smaller deletion (3.2 kbp) to yield AD169*var*UC and then underwent a larger deletion to yield AD169*var*UK and AD169*var*ATCC or whether AD169varUC represents an alternative passage to the other variants that lost 3.2 kbp during cell culture. Another possibility is that AD169*var*UC is a mixture, a proportion of which contains this additional segment at the right end of U_L, and the remainder contains the large deletion found in AD169*var*UK and AD169*var*ATCC. PCR using primers designed either side of the 15 kbp deletion could be performed to investigate this possibility.

References

Adler, S. P., Plotkin, S. A., Gonczol, E., Cadoz, M., Meric, C., Wang, J. B., Dellamonica, P., Best, A. M., Zahradnik, J., Pincus, S., Berencsi, K., Cox, W. I. and Gyulai, Z. (1999). A canarypox vector expressing cytomegalovirus (CMV) glycoprotein B primes for antibody responses to a live attenuated CMV vaccine (Towne). J Infect Dis 180, 843-6.

Adler, S. P. (1995). Immunoprophylaxis against cytomegalovirus disease. Scand J Infect Dis Suppl 99, 105-9.

Ahlfors, K., Ivarsson, S. A., Johnsson, T. and Svanberg, L. (1982). Primary and secondary maternal cytomegalovirus infections and their relation to congenital infection. Analysis of maternal sera. Acta Paediatr Scand 71, 109-13.

Ahuja, S. K., Gao, J. L. and Murphy, P. M. (1994). Chemokine receptors and molecular mimicry. Immunol Today 15, 281-7.

Akter, P., Cunningham, C., McSharry, B. P., Dolan, A., Addison, C., Dargan, D. J., Hassan-Walker, A. F., Emery, V. C., Griffiths, P. D., Wilkinson, G. W. and Davison, A. J. (2003). Two novel spliced genes in human cytomegalovirus. J Gen Virol 84, 1117-22.

Alderete, J. P., Jarrahian, S. and Geballe, A. P. (1999). Translational effects of mutations and polymorphisms in a repressive upstream open reading frame of the human cytomegalovirus UL4 gene. J Virol 73, 8330-7.

Aoki, T., Hirono, I., Kurokawa, K., Fukuda, H., Nahary, R., Eldar, A., Davison, A. J., Waltzek, T. B., Bercovier, H. and Hedrick, R. P. (2007). Genome sequences of three koi herpesvirus isolates representing the expanding distribution of an emerging disease threatening koi and common carp worldwide. J Virol 81, 5058-65.

Aquino, V. H. and Figueiredo, L. T. (2000). High prevalence of renal transplant recipients infected with more than one cytomegalovirus glycoprotein B genotype. J Med Virol 61, 138-42.

Arav-Boger, R., Willoughby, R. E., Pass, R. F., Zong, J. C., Jang, W. J., Alcendor, D. and Hayward, G. S. (2002). Polymorphisms of the cytomegalovirus (CMV)-encoded tumor necrosis factor-alpha and beta-chemokine receptors in congenital CMV disease. J Infect Dis 186, 1057-64.

Arav-Boger, R., Zong, J. C. and Foster, C. B. (2005). Loss of linkage disequilibrium and accelerated protein divergence in duplicated cytomegalovirus chemokine genes. Virus Genes 31, 65-72.

Arav-Boger, R., Foster, C. B., Zong, J. C. and Pass, R. F. (2006). Human cytomegalovirus-encoded alpha -chemokines exhibit high sequence variability in congenitally infected newborns. J Infect Dis 193, 788-91.

Arav-Boger, R., Battaglia, C. A., Lazzarotto, T., Gabrielli, L., Zong, J. C., Hayward, G. S., Diener-West, M. and Landini, M. P. (2006a). Cytomegalovirus (CMV)-encoded UL144 (truncated tumor necrosis factor receptor) and outcome of congenital CMV infection. J Infect Dis 194, 464-73.

Atalay, R., Zimmermann, A., Wagner, M., Borst, E., Benz, C., Messerle, M. and Hengel, H. (2002). Identification and expression of human cytomegalovirus transcription units coding for two distinct Fc gamma receptor homologs. J Virol 76, 8596-608.

Azad, R. F., Driver, V. B., Tanaka, K., Crooke, R. M. and Anderson, K. P. (1993). Antiviral activity of a phosphorothioate oligonucleotide complementary to RNA of the human cytomegalovirus major immediate-early region. Antimicrob Agents Chemother 37, 1945-54.

Baggiolini, M., Loetscher, P. and Moser, B. (1995). Interleukin-8 and the chemokine family. Int J Immunopharmacol 17, 103-8.

Baggiolini, M. (1995a). Activation and recruitment of neutrophil leukocytes. Clin Exp Immunol 101 Suppl 1, 5-6.

Baggiolini, M., Dewald, B. and Moser, B. (1997). Human chemokines: an update. Annu Rev Immunol 15, 675-705.

Baggiolini, M. and Loetscher, P. (2000). Chemokines in inflammation and immunity. Immunol Today 21, 418-20.

Baldanti, F., Lurain, N. and Gerna, G. (2004). Clinical and biologic aspects of human cytomegalovirus resistance to antiviral drugs. Hum Immunol 65, 403-9.

Baldick, C. J., Jr. and Shenk, T. (1996). Proteins associated with purified human cytomegalovirus particles. J Virol 70, 6097-105.

Baldwin, E. T., Weber, I. T., St Charles, R., Xuan, J. C., Appella, E., Yamada, M., Matsushima, K., Edwards, B. F., Clore, G. M., Gronenborn, A. M. and et al. (1991). Crystal structure of interleukin 8: symbiosis of NMR and crystallography. Proc Natl Acad Sci U S A 88, 502-6.

Bale, J. F., Jr., Zimmerman, B., Dawson, J. D., Souza, I. E., Petheram, S. J. and Murph, J. R. (1999). Cytomegalovirus transmission in child care homes. Arch Pediatr Adolesc Med 153, 75-9.

Bale, J. F., Jr., Petheram, S. J., Robertson, M., Murph, J. R. and Demmler, G. (2001). Human cytomegalovirus *a* sequence and UL144 variability in strains from infected children. J Med Virol 65, 90-6.

Bar, M., Shannon-Lowe, C. and Geballe, A. P. (2001). Differentiation of human cytomegalovirus genotypes in immunocompromised patients on the basis of UL4 gene polymorphisms. J Infect Dis 183, 218-225.

Bechtel, J. T. and Shenk, T. (2002). Human cytomegalovirus UL47 tegument protein functions after entry and before immediate-early gene expression. J Virol 76, 1043-50.

Bego, M., Maciejewski, J., Khaiboullina, S., Pari, G. and St Jeor, S. (2005). Characterization of an antisense transcript spanning the UL81-82 locus of human cytomegalovirus. J Virol 79, 11022-34.

Benedict, C. A., Butrovich, K. D., Lurain, N. S., Corbeil, J., Rooney, I., Schneider, P., Tschopp, J. and Ware, C. F. (1999). Cutting edge: a novel viral

TNF receptor superfamily member in virulent strains of human cytomegalovirus. J Immunol 162, 6967-70.

Berencsi, K., Gyulai, Z., Gonczol, E., Pincus, S., Cox, W. I., Michelson, S., Kari, L., Meric, C., Cadoz, M., Zahradnik, J., Starr, S. and Plotkin, S. (2001). A canarypox vector-expressing cytomegalovirus (CMV) phosphoprotein 65 induces long-lasting cytotoxic T cell responses in human CMV-seronegative subjects. J Infect Dis 183, 1171-9.

Bernstein, D. I., Schleiss, M. R., Berencsi, K., Gonczol, E., Dickey, M., Khoury, P., Cadoz, M., Meric, C., Zahradnik, J., Duliege, A. M. and Plotkin, S. (2002). Effect of previous or simultaneous immunization with canarypox expressing cytomegalovirus (CMV) glycoprotein B (gB) on response to subunit gB vaccine plus MF59 in healthy CMV-seronegative adults. J Infect Dis 185, 686-90.

Berti R., Soldan S., Akhyani N., McFarland H., Jacobson S. (2000). Extended observations on the association of HHV-6 and multiplesclerosis. J Neurovirol 6, 85–87.

Beyari, M. M., Hodgson, T. A., Kondowe, W., Molyneux, E. M., Scully, C., Porter, S. R. and Teo, C. G. (2005). Inter- and intra-person cytomegalovirus infection in Malawian families. J Med Virol 75, 575-82.

Bhella, D., Rixon, F. J. and Dargan, D. J. (2000). Cryomicroscopy of human cytomegalovirus virions reveals more densely packed genomic DNA than in herpes simplex virus type 1. J Mol Biol 295, 155-61.

Biron, C. A., Byron, K. S. and Sullivan, J. L. (1989). Severe herpesvirus infections in an adolescent without natural killer cells. N Engl J Med 320, 1731-5.

Bogner, E., Radsak, K. and Stinski, M. F. (1998). The gene product of human cytomegalovirus open reading frame UL56 binds the pac motif and has specific nuclease activity. J Virol 72, 2259-64.

Bolovan-Fritts, C. A., Mocarski, E. S. and Wiedeman, J. A. (1999). Peripheral blood CD14(+) cells from healthy subjects carry a circular conformation of latent cytomegalovirus genome. Blood 93, 394-8.

Boppana, S. B., Pass, R. F., Britt, W. J., Stagno, S. and Alford, C. A. (1992). Symptomatic congenital cytomegalovirus infection: neonatal morbidity and mortality. Pediatr Infect Dis J 11, 93-9.

Boppana, S. B., Rivera, L. B., Fowler, K. B., Mach, M. and Britt, W. J. (2001). Intrauterine transmission of cytomegalovirus to infants of women with preconceptional immunity. N Engl J Med 344, 1366-71.

Borghi, E., Pagani, E., Mancuso, R., Delbue, S., Valli, M., Mazziotti, R., Giordano, L., Micheli, R. and Ferrante, P. (2005). Detection of herpesvirus-6A in a case of subacute cerebellitis and myoclonic dystonia. J Med Virol 75, 427-9.

Britt, W. J. and Vugler, L. G. (1989). Processing of the gp55-116 envelope glycoprotein complex (gB) of human cytomegalovirus. J Virol 63, 403-10.

Britt, W. J. and Boppana, S. (2004). Human cytomegalovirus virion proteins. Hum Immunol 65, 395-402.

Brown, J. M., Kaneshima, H. and Mocarski, E. S. (1995). Dramatic interstrain differences in the replication of human cytomegalovirus in SCID-hu mice. J Infect Dis 171, 1599-603.

Brytting, M., Wahlberg, J., Lundeberg, J., Wahren, B., Uhlen, M. and Sundqvist, V. A. (1992). Variations in the cytomegalovirus major immediate-early gene found by direct genomic sequencing. J Clin Microbiol 30, 955-60.

Butcher, S. J., Aitken, J., Mitchell, J., Gowen, B. and Dargan, D. J. (1998). Structure of the human cytomegalovirus B capsid by electron cryomicroscopy and image reconstruction. J Struct Biol 124, 70-6.

Cantrell, S. R. and Bresnahan, W. A. (2006). Human cytomegalovirus (HCMV) UL82 gene product (pp71) relieves hDaxx-mediated repression of HCMV replication. J Virol 80, 6188-91.

Carraro, E. and Granato, C. F. (2003). Single human cytomegalovirus gB genotype shed in multiple sites at the time of diagnosis in renal transplant recipients. J Med Virol 70, 240-3.

Carrigan D.R., Drobyski W.R., Russler S.K., Tapper M.A., Knox K.K., Ash R.C. (1991). Interstitial pneumonitis associated with human herpesvirus-6 infection after marrow transplantation. Lancet. 338(8760), 147-9.

Cha, T. A., Tom, E., Kemble, G. W., Duke, G. M., Mocarski, E. S. and 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., Guo, H., Jiang, Y., Wan, J. S., Bittner, A., Frueh, K., Jackson, M. R., Peterson, P. A., Erlander, M. G. and Ghazal, P. (1999). DNA microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression. J Virol 73, 5757-66.

Chang, C. P., Vesole, D. H., Nelson, J., Oldstone, M. B. and Stinski, M. F. (1989). Identification and expression of a human cytomegalovirus early glycoprotein. J Virol 63, 3330-7.

Chang J., Schmid M., Rixon F., and Chiu W. (2007). Electron Cryotomography Reveals the Portal in the Herpesvirus Capsid. J Virol 81, 2065-2068.

Chee, M. S., Bankier, A. T., Beck, S., Bohni, R., Brown, C. M., Cerny, R., Horsnell, T., Hutchison, C. A., 3rd, Kouzarides, T., Martignetti, J. A. and et al. (1990). Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. Curr Top Microbiol Immunol 154, 125-69.

Chen D., Jiang H., Lee M., Liu F., and Zhou Z. (1999). Three-dimensional visualization of tegument/capsid interactions in the intact human cytomegalovirus. Virology. 260(1), 10-6.

Chen, S. F., Tu, W. W., Sharp, M. A., Tongson, E. C., He, X. S., Greenberg, H. B., Holmes, T. H., Wang, Z., Kemble, G., Manganello, A. M., Adler, S. P., Dekker, C. L., Lewis, D. B. and Arvin, A. M. (2004). Antiviral CD8 T cells in the control of primary human cytomegalovirus infection in early childhood. J Infect Dis 189, 1619-27.

Chen, F. H., Samson, K. T., Chen, H., Pan, S. N., He, Z. X., Iikura, Y. and Shioda, S. (2004a). Clinical applications of real-time PCR for diagnosis and treatment of human cytomegalovirus infection in children. Pediatr Allergy Immunol 15, 210-5.

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. Embo J 5, 823-6.

Chou, S. W. and Dennison, K. M. (1991). Analysis of interstrain variation in cytomegalovirus glycoprotein B sequences encoding neutralization-related epitopes. J Infect Dis 163, 1229-34.

Chou, S. (1992). Molecular epidemiology of envelope glycoprotein H of human cytomegalovirus. J Infect Dis 166, 604-7.

Chou, S., Lurain, N. S., Thompson, K. D., Miner, R. C. and Drew, W. L. (2003). Viral DNA polymerase mutations associated with drug resistance in human cytomegalovirus. J Infect Dis 188, 32-9.

Cicin-Sain, L., Podlech, J., Messerle, M., Reddehase, M. J. and Koszinowski, U. H. (2005). Frequent coinfection of cells explains functional in vivo complementation between cytomegalovirus variants in the multiply infected host. J Virol 79, 9492-502.

Clark-Lewis, I., Schumacher, C., Baggiolini, M. and Moser, B. (1991). Structure-activity relationships of interleukin-8 determined using chemically synthesized analogs. Critical role of NH2-terminal residues and evidence for uncoupling of neutrophil chemotaxis, exocytosis, and receptor binding activities. J Biol Chem 266, 23128-34.

Clark-Lewis, I., Kim, K. S., Rajarathnam, K., Gong, J. H., Dewald, B., Moser, B., Baggiolini, M. and Sykes, B. D. (1995). Structure-activity relationships of chemokines. J Leukoc Biol 57, 703-11.

Clore, G. M., Appella, E., Yamada, M., Matsushima, K. and Gronenborn, A. M. (1990). Three-dimensional structure of interleukin 8 in solution. Biochemistry 29, 1689-96.

Compton T. (2004) Receptors and immune sensors: the complex entry path of human cytomegalovirus. Trends in Cell Biology. 14 (1), 5-8.

Coaquette, A., Bourgeois, A., Dirand, C., Varin, A., Chen, W. and Herbein, G. (2004). Mixed cytomegalovirus glycoprotein B genotypes in immunocompromised patients. Clin Infect Dis 39, 155-61.

Craig, J. M., Macauley, J. C., Weller, T. H. and Wirth, P. (1957). Isolation of intranuclear inclusion producing agents from infants with illnesses resembling cytomegalic inclusion disease. Proc Soc Exp Biol Med 94, 4-12.

Creighton, T. (1993). Proteins: Structures and Molecular Properties, 2nd edition edn. New York: W.H. Freeman and Co.

Dal Monte, P., Pignatelli, S., Rossini, G. and Landini, M. P. (2004). Genomic variants among human cytomegalovirus (HCMV) clinical isolates: the glycoprotein n (gN) paradigm. Hum Immunol 65, 387-94.

Dargan, D. J., Jamieson, F. E., MacLean, J., Dolan, A., Addison, C. and McGeoch, D. J. (1997). The published DNA sequence of human cytomegalovirus strain AD169 lacks 929 base pairs affecting genes UL42 and UL43. J Virol 71, 9833-6.

Davis, C. L., Field, D., Metzgar, D., Saiz, R., Morin, P. A., Smith, I. L., Spector, S. A. and Wills, C. (1999). Numerous length polymorphisms at short tandem repeats in human cytomegalovirus. J Virol 73, 6265-70.

Davison, A. J., Dargan, D. J. and Stow, N. D. (2002). Fundamental and accessory systems in herpesviruses. Antiviral Res 56, 1-11.

Davison, A. J. (2002a). Evolution of the herpesviruses. Vet Microbiol 86, 69-88.

Davison, A. (2002b). Comments on the phylogenetics and evolution of herpesviruses and other large DNA viruses. Virus Res 82, 127-32.

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

Davison, A. J., Akter, P., Cunningham, C., Dolan, A., Addison, C., Dargan, D. J., Hassan-Walker, A. F., Emery, V. C., Griffiths, P. D. and Wilkinson, G. W. (2003a). Homology between the human cytomegalovirus RL11 gene family and human adenovirus E3 genes. J Gen Virol 84, 657-63.

Davison, A. J. and Stow, N. D. (2005). New genes from old: redeployment of dUTPase by herpesviruses. J Virol 79, 12880-92.

Davison, A.J., R. Eberle, GS Hayward, DJ McGeoch, AC Minson, PE Pellet, B Roizman, MJ Studdert and E Thiry. (2005a). Herpesviridae. Virus taxonomy: 8th report of the international committee on Taxonomy of viruses. 193-212. Elsevier/Academic press, London, England.

Day EK, Carmichael AJ, ten Berge IJ, Waller EC, Sissons JG, Wills MR. (2007) Rapid CD8+ T cell repertoire focusing and selection of high-affinity clones into memory following primary infection with a persistent human virus: human cytomegalovirus. J Immunol. 179(5), 3203-13.

Dedicoat, M. and Newton, R. (2003). Review of the distribution of Kaposi's sarcoma-associated herpesvirus (KSHV) in Africa in relation to the incidence of Kaposi's sarcoma. Br J Cancer 88, 1-3.

Dedicoat, M., Newton, R., Alkharsah, K. R., Sheldon, J., Szabados, I., Ndlovu, B., Page, T., Casabonne, D., Gilks, C. F., Cassol, S. A., Whitby, D. and Schulz, T. F. (2004). Mother-to-child transmission of human herpesvirus-8 in South Africa. J Infect Dis 190, 1068-75.

Dewhurst, S., Skrincosky, D. and van Loon, N. (1997). Human herpesvirus 6. Expert Rev Mol Med 1997, 1-17.

Dohner, D. E., Adams, S. G. and Gelb, L. D. (1988). Recombination in tissue culture between varicella-zoster virus strains. J Med Virol 24, 329-41.

Dohner, K. and Sodeik, B. (2005). The role of the cytoskeleton during viral infection. Curr Top Microbiol Immunol 285, 67-108.

Dohner, K., Nagel, C. H. and Sodeik, B. (2005a). Viral stop-and-go along microtubules: taking a ride with dynein and kinesins. Trends Microbiol 13, 320-7.

Dolan, A., Cunningham, C., Hector, R. D., Hassan-Walker, A. F., Lee, L., Addison, C., Dargan, D. J., McGeoch, D. J., Gatherer, D., Emery, V. C., Griffiths, P. D., Sinzger, C., McSharry, B. P., Wilkinson, G. W. and Davison, A. J. (2004). Genetic content of wild-type human cytomegalovirus. J Gen Virol 85, 1301-12.

Drew, W. L. (1988). Diagnosis of cytomegalovirus infection. Rev Infect Dis 10 Suppl 3, S468-76.

Dunn, W., Chou, C., Li, H., Hai, R., Patterson, D., Stolc, V., Zhu, H. and Liu, F. (2003). Functional profiling of a human cytomegalovirus genome. Proc Natl Acad Sci U S A 100, 14223-8.

Efstathiou, S. and Preston, C. M. (2005). Towards an understanding of the molecular basis of herpes simplex virus latency. Virus Res 111, 108-19.

Elek, S. D. and Stern, H. (1974). Development of a vaccine against mental retardation caused by cytomegalovirus infection in utero. Lancet 1, 1-5.

Elkington R, Walker S, Crough T, Menzies M, Tellam J, Bharadwaj M, Khanna R. (2003). Ex Vivo Profiling of CD8$^+$-T-Cell Responses to Human Cytomegalovirus Reveals Broad and Multispecific Reactivities in Healthy Virus Carriers. J Virol. 77(9), 5226-40.

Eriksson, B., Oberg, B. and Wahren, B. (1982). Pyrophosphate analogues as inhibitors of DNA polymerases of cytomegalovirus, herpes simplex virus and cellular origin. Biochim Biophys Acta 696, 115-23.

Faqi, A. S., Klug, A., Merker, H. J. and Chahoud, I. (1997). Ganciclovir induces reproductive hazards in male rats after short-term exposure. Hum Exp Toxicol 16, 505-11.

Felsenstein, J. (1989). PHYLIP: phylogeny inference package (version 3.2). Cladistics 5, 164-166.

Field, D and C. Wills. (1998). Abundant microsatellite polymorphism in S. cerevisiae, and the different distributions of microsatellites in prokaryotes and eukaryotes, result from strong mutation pressures and a variety of selective forces. PNAS. 95:1647-52.

Fischer, G. F., Majdic, O., Gadd, S. and Knapp, W. (1990). Signal transduction in lymphocytic and myeloid cells via CD24, a new member of phosphoinositol-anchored membrane molecules. J Immunol 144, 638-41.

Fraile-Ramos A., Pelchen-Matthews A., Kledal T., Browne H., Schwartz T., Marsh M. (2002). Localization of HCMV UL33 and US27 in endocytic compartments and viral membranes. Traffic. 3(3), 218-32.

Fries, B. C., Chou, S., Boeckh, M. and Torok-Storb, B. (1994). Frequency distribution of cytomegalovirus envelope glycoprotein genotypes in bone marrow transplant recipients. J Infect Dis 169, 769-74.

Gandhi, M. K. and Khanna, R. (2004). Human cytomegalovirus: clinical aspects, immune regulation, and emerging treatments. Lancet Infect Dis 4, 725-38.

Gerna, G., Baldanti, F., Zavattoni, M., Sarasini, A., Percivalle, E. and Revello, M. G. (1992). Monitoring of ganciclovir sensitivity of multiple human cytomegalovirus strains coinfecting blood of an AIDS patient by an immediate-early antigen plaque assay. Antiviral Res 19, 333-45.

Gerna, G., Zavattoni, M., Percivalle, E., Grossi, P., Torsellini, M. and Revello, M. G. (1998). Rising levels of human cytomegalovirus (HCMV) antigenemia during initial antiviral treatment of solid-organ transplant recipients with primary HCMV infection. J Clin Microbiol 36, 1113-6.

Gerna, G., Percivalle, E., Lilleri, D., Lozza, L., Fornara, C., Hahn, G., Baldanti, F. and Revello, M. G. (2005). Dendritic-cell infection by human cytomegalovirus is restricted to strains carrying functional UL131-128 genes and mediates efficient viral antigen presentation to CD8+ T cells. J Gen Virol 86, 275-84.

Gershon, A. A., Mervish, N., LaRussa, P., Steinberg, S., Lo, S. H., Hodes, D., Fikrig, S., Bonagura, V. and Bakshi, S. (1997). Varicella-zoster virus infection in children with underlying human immunodeficiency virus infection. J Infect Dis 176, 1496-500.

Gibson, W. (1996). Structure and assembly of the virion. Intervirology 39, 389-400.

Giesen, K., Radsak, K. and Bogner, E. (2000). The potential terminase subunit of human cytomegalovirus, pUL56, is translocated into the nucleus by its own nuclear localization signal and interacts with importin alpha. J Gen Virol 81, 2231-44.

Gleaves, C. A., Smith, T. F., Shuster, E. A. and Pearson, G. R. (1984). Rapid detection of cytomegalovirus in MRC-5 cells inoculated with urine specimens by using low-speed centrifugation and monoclonal antibody to an early antigen. J Clin Microbiol 19, 917-9.

Goldmacher, V. S., Bartle, L. M., Skaletskaya, A., Dionne, C. A., Kedersha, N. L., Vater, C. A., Han, J. W., Lutz, R. J., Watanabe, S., Cahir McFarland, E. D., Kieff, E. D., Mocarski, E. S. and Chittenden, T. (1999). A cytomegalovirus-encoded mitochondria-localized inhibitor of apoptosis structurally unrelated to Bcl-2. Proc Natl Acad Sci U S A 96, 12536-41.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11, 725-36.

Gompels U.A., Nicholas J., Lawrence G., Jones M., Thomson B.J., Martin M.E., Efstathiou S., Craxton M., Macaulay H.A. (1995). The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution. Virology. 209(1), 29-51.

Goris, A., Maranian, M., Walton, A., Yeo, T. W., Ban, M., Gray, J., Dubois, B., Compston, A. and Sawcer, S. (2006). CD24 Ala/Val polymorphism and multiple sclerosis. J Neuroimmunol 175, 200-2.

Gorman, S., Harvey, N. L., Moro, D., Lloyd, M. L., Voigt, V., Smith, L. M., Lawson, M. A. and Shellam, G. R. (2006). Mixed infection with multiple strains of murine cytomegalovirus occurs following simultaneous or sequential infection of immunocompetent mice. J Gen Virol 87, 1123-32.

Gray, W. L., Mullis, L. B. and Soike, K. F. (2001). Expression of the simian varicella virus glycoprotein E. Virus Res 79, 27-37.

Gretch, D. R., Kari, B., Rasmussen, L., Gehrz, R. C. and Stinski, M. F. (1988). Identification and characterization of three distinct families of glycoprotein complexes in the envelopes of human cytomegalovirus. J Virol 62, 875-81.

Gupta, N.K., Ohtsuka,E., Sgaramella,V., Buchi,H., Kumar,A., Weber,H. and Khorana,H.G. (1968) Studies on polynucleotides, 88. Enzymatic joining of chemically synthesized segments corresponding to the gene for alanine-tRNA. Proc. Natl Acad. Sci. USA, 60, 1338–1344.

Hahn, G., Revello, M. G., Patrone, M., Percivalle, E., Campanini, G., Sarasini, A., Wagner, M., Gallina, A., Milanesi, G., Koszinowski, U., Baldanti, F. and Gerna, G. (2004). Human cytomegalovirus UL131-128 genes are indispensable for virus growth in endothelial cells and virus transfer to leukocytes. J Virol 78, 10023-33.

Hassan-Walker, A. F., Okwuadi, S., Lee, L., Griffiths, P. D. and Emery, V. C. (2004). Sequence variability of the alpha-chemokine UL146 from clinical strains of human cytomegalovirus. J Med Virol 74, 573-9.

Hayajneh, W. A., Colberg-Poley, A. M., Skaletskaya, A., Bartle, L. M., Lesperance, M. M., Contopoulos-Ioannidis, D. G., Kedersha, N. L. and Goldmacher, V. S. (2001). The sequence and antiapoptotic functional domains of the human cytomegalovirus UL37 exon 1 immediate early protein are conserved in multiple primary strains. Virology 279, 233-40.

Hayajneh, W. A., Contopoulos-Ioannidis, D. G., Lesperance, M. M., Venegas, A. M. and Colberg-Poley, A. M. (2001a). The carboxyl terminus of the human cytomegalovirus UL37 immediate-early glycoprotein is conserved in primary strains and is important for transactivation. J Gen Virol 82, 1569-79.

He, R., Ruan, Q., Qi, Y., Ma, Y. P., Huang, Y. J., Sun, Z. R. and Ji, Y. H. (2006). Sequence variability of human cytomegalovirus UL146 and UL147 genes in low-passage clinical isolates. Intervirology 49, 215-23.

Heineman, T. C., Schleiss, M., Bernstein, D. I., Spaete, R. R., Yan, L., Duke, G., Prichard, M., Wang, Z., Yan, Q., Sharp, M. A., Klein, N., Arvin, A. M. and Kemble, G. (2006). A phase 1 study of 4 live, recombinant human cytomegalovirus Towne/Toledo chimeric vaccines. J Infect Dis 193, 1350-60.

Henderson, L. M., Katz, J. B., Erickson, G. A. and Mayfield, J. E. (1990). In vivo and in vitro genetic recombination between conventional and gene-deleted vaccine strains of pseudorabies virus. Am J Vet Res 51, 1656-62.

Henniker, A. J. (2001). Cd24. J Biol Regul Homeost Agents 15, 182-4.

Hey, J. (2005). On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biol 3, 193.

Hitomi, S., Kozuka-Hata, H., Chen, Z., Sugano, S., Yamaguchi, N. and Watanabe, S. (1997). Human cytomegalovirus open reading frame UL11 encodes a highly polymorphic protein expressed on the infected cell surface. Arch Virol 142, 1407-27.

Hobom, U., Brune, W., Messerle, M., Hahn, G. and 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, 7720-9.

Homa, F. L. and Brown, J. C. (1997). Capsid assembly and DNA packaging in herpes simplex virus. Rev Med Virol 7, 107-122.

Homman-Loudiyi, M., Hultenby, K., Britt, W. and Soderberg-Naucler, C. (2003). Envelopment of human cytomegalovirus occurs by budding into Golgi-derived vacuole compartments positive for gB, Rab 3, trans-golgi network 46, and mannosidase II. J Virol 77, 3191-203.

Honess, R. W. (1984). Herpes simplex and 'the herpes complex': diverse observations and a unifying hypothesis. The eighth Fleming lecture. J Gen Virol 65 (Pt 12), 2077-107.

Huber, M. T. and Compton, T. (1998). The human cytomegalovirus UL74 gene encodes the third component of the glycoprotein H-glycoprotein L-containing envelope complex. J Virol 72, 8191-7.

Humar, A., Kumar, D., Gilbert, C. and Boivin, G. (2003). Cytomegalovirus (CMV) glycoprotein B genotypes and response to antiviral therapy, in solid-organ-transplant recipients with CMV disease. J Infect Dis 188, 581-4.

Hurst, LD (2002) The Ka/Ks ratio: Diagnosing the form of sequence evolution. Trends Genet 18, 486.

Hutt-Fletcher, L. M. (2007). Epstein-Barr virus entry. J Virol 81, 7825-32.

Ishov, A. M., Vladimirova, O. V. and Maul, G. G. (2002). Daxx-mediated accumulation of human cytomegalovirus tegument protein pp71 at ND10 facilitates initiation of viral infection at these nuclear domains. J Virol 76, 7705-12.

Iversen, A. C., Norris, P. S., Ware, C. F. and Benedict, C. A. (2005). Human NK cells inhibit cytomegalovirus replication through a noncytolytic mechanism involving lymphotoxin-dependent induction of IFN-beta. J Immunol 175, 7568-74.

Jarvis M.A., Nelson J.A. (2007). Human cytomegalovirus tropism for endothelial cells: not all endothelial cells are created equal. J Virol. 81(5), 2095-101.

Jenkins, C., Abendroth, A. and Slobedman, B. (2004). A novel viral transcript with homology to human interleukin-10 is expressed during latent human cytomegalovirus infection. J Virol 78, 1440-7.

Judo, M. S., Wedel, A. B. and Wilson, C. (1998). Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res 26, 1819-25.

Kalejta, R. F. and Shenk, T. (2003). The human cytomegalovirus UL82 gene product (pp71) accelerates progression through the G1 phase of the cell cycle. J Virol 77, 3451-9.

Kall, L., Krogh, A. and Sonnhammer, E. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res 35, W429-32.

Kari, B. and Gehrz, R. (1990). Analysis of human antibody responses to human cytomegalovirus envelope glycoproteins found in two families of disulfide linked glycoprotein complexes designated gC-I and gC-II. Arch Virol 114, 213-28.

Karran, L., Jones, M., Morley, G., van Noorden, S., Smith, P., Lampert, I. and Griffin, B. E. (1995). Expression of a B-cell marker, CD24, on nasopharyngeal carcinoma cells. Int J Cancer 60, 562-6.

Katoh, K., Kuma, K., Toh, H. and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33, 511-8.

Khanna, R. and Diamond, D. J. (2006). Human cytomegalovirus vaccine: time to look for alternative options. Trends Mol Med 12, 26-33.

Kimberlin DW, Lin CY, Sánchez PJ, Demmler GJ, Dankner W, Shelton M, Jacobs RF, Vaudry W, Pass RF, Kiell JM, Soong SJ, Whitley RJ; National Institute of Allergy and Infectious Diseases Collaborative Antiviral Study Group. **(2003). Effect of ganciclovir therapy on hearing in symptomatic congenital cytomegalovirus disease involving the central nervous system: a randomized,** controlled trial. J Pediatr. 143(1), 16-25.

Kovacs A., Schluchter M., Easley K., Demmler G., Shearer W., La Russa P., Pitt J., Cooper E., Goldfarb J., Hodes D., Kattan M. and McIntosh K. (1999) Cytomegalovirus infection and HIV-1 disease progression in infants born to HIV-1-infected women. Pediatric Pulmonary and Cardiovascular Complications of Vertically Transmitted HIV Infection Study Group. N Engl J Med. 341(19), 1476-7.

Kondo K, Kondo T, Okuno T, Takahashi M and Yamanishi K. (1991). Latent human herpesvirus 6 infection of human monocytes/macrophages. J Gen Virol 72, 1401-08.

Kondo, K., Kaneshima, H. and Mocarski, E. S. (1994). Human cytomegalovirus latent infection of granulocyte-macrophage progenitors. Proc Natl Acad Sci U S A 91, 11879-83.

Kondo, K. and Mocarski, E. S. (1995). Cytomegalovirus latency and latency-specific transcription in hematopoietic progenitors. Scand J Infect Dis Suppl 99, 63-7.

Kosuge, H. (2000). HHV-6, 7 and their related diseases. J Dermatol Sci 22, 205-12.

Kotenko, S. V., Saccani, S., Izotova, L. S., Mirochnitchenko, O. V. and Pestka, S. (2000). Human cytomegalovirus harbors its own unique IL-10 homolog (cmvIL-10). Proc Natl Acad Sci U S A 97, 1695-700.

Kristiansen, G., Sammar, M. and Altevogt, P. (2004). Tumour biological aspects of CD24, a mucin-like adhesion molecule. J Mol Histol 35, 255-62.

Krosky, P. M., Underwood, M. R., Turk, S. R., Feng, K. W., Jain, R. K., Ptak, R. G., Westerman, A. C., Biron, K. K., Townsend, L. B. and Drach, J. C. (1998). Resistance of human cytomegalovirus to benzimidazole ribonucleosides maps to two open reading frames: UL89 and UL56. J Virol 72, 4721-8.

Kumar, S., Tamura, K. and Nei, M. (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5, 150-63.

Kutok, J. L. and Wang, F. (2006). Spectrum of Epstein-Barr virus-associated diseases. Annu Rev Pathol 1, 375-404.

Leong, S. R., Lowman, H. B., Liu, J., Shire, S., Deforge, L. E., Gillece-Castro, B. L., McDowell, R. and Hebert, C. A. (1997). IL-8 single-chain homodimers and heterodimers: interactions with chemokine receptors CXCR1, CXCR2, and DARC. Protein Sci 6, 609-17.

Lukacsi, A., Tarodi, B., Endreffy, E., Babinszki, A., Pal, A. and Pusztai, R. (2001). Human cytomegalovirus gB genotype 1 is dominant in congenital infections in South Hungary. J Med Virol 65, 537-42.

Lurain, N. S., Kapell, K. S., Huang, D. D., Short, J. A., Paintsil, J., Winkfield, E., Benedict, C. A., Ware, C. F. and Bremer, J. W. (1999). Human cytomegalovirus UL144 open reading frame: sequence hypervariability in low-passage clinical isolates. J Virol 73, 10040-50.

Lurain, N. S., Bhorade, S. M., Pursell, K. J., Avery, R. K., Yeldandi, V. V., Isada, C. M., Robert, E. S., Kohn, D. J., Arens, M. Q., Garrity, E. R., Taege, A. J., Mullen, M. G., Todd, K. M., Bremer, J. W. and Yen-Lieberman, B. (2002). Analysis and characterization of antiviral drug-resistant cytomegalovirus isolates from solid organ transplant recipients. J Infect Dis 186, 760-8.

Lurain, N. S., Fox, A. M., Lichy, H. M., Bhorade, S. M., Ware, C. F., Huang, D. D., Kwan, S. P., Garrity, E. R. and Chou, S. (2006). Analysis of the human cytomegalovirus genomic region from UL146 through UL147A reveals sequence hypervariability, genotypic stability, and overlapping transcripts. Virol J 3, 4.

Malm, G. and Engman, M. L. (2007). Congenital cytomegalovirus infections. Semin Fetal Neonatal Med 12, 154-9.

Mao, Z. Q., He, R., Sun, M., Qi, Y., Huang, Y. J. and Ruan, Q. (2007). The relationship between polymorphisms of HCMV UL144 ORF and clinical manifestations in 73 strains with congenital and/or perinatal HCMV infection. Arch Virol 152, 115-24.

Mattick, C., Dewin, D., Polley, S., Sevilla-Reyes, E., Pignatelli, S., Rawlinson, W., Wilkinson, G., Dal Monte, P. and Gompels, U. A. (2004). Linkage of human cytomegalovirus glycoprotein gO variant groups identified from worldwide clinical isolates with gN genotypes, implications for disease associations and evidence for N-terminal sites of positive selection. Virology 318, 582-97.

Mayo, K. H., Roongta, V., Ilyina, E., Milius, R., Barker, S., Quinlan, C., La Rosa, G. and Daly, T. J. (1995). NMR solution structure of the 32-kDa platelet factor 4 ELR-motif N-terminal chimera: a symmetric tetramer. Biochemistry 34, 11399-409.

Mayr, E. (2001). What evolution is. Basic Books. New York.

McGeoch, D. J., Dolan, A. and Ralph, A. C. (2000). Toward a comprehensive phylogeny for mammalian and avian herpesviruses. J Virol 74, 10401-6.

McGeoch, D. J., Gatherer, D. and Dolan, A. (2005). On phylogenetic relationships among major lineages of the Gammaherpesvirinae. J Gen Virol 86, 307-16.

McGeoch, D. J., Rixon, F. J. and Davison, A. J. (2006). Topics in herpesvirus genomics and evolution. Virus Res 117, 90-104.

Mendelson, M., Monard, S., Sissons, P. and Sinclair, J. (1996). Detection of endogenous human cytomegalovirus in CD34+ bone marrow progenitors. J Gen Virol 77 (Pt 12), 3099-102.

Mettenleiter, T. C. (2004). Budding events in herpesvirus morphogenesis. Virus Res 106, 167-80.

Meyer-Konig, U., Schrage, B., Huzly, D., Bongarts, A. and Hufert, F. T. (1998). High variability of cytomegalovirus glycoprotein B gene and frequent multiple infections in HIV-infected patients with low CD4 T-cell count. Aids 12, 2228-30.

Meyer-Konig, U., Ebert, K., Schrage, B., Pollak, S. and Hufert, F. T. (1998a). Simultaneous infection of healthy people with multiple human cytomegalovirus strains. Lancet 352, 1280-1.

Miller-Kittrell, M., Sai, J., Penfold, M., Richmond, A. and Sparer, T. E. (2007). Functional characterization of chimpanzee cytomegalovirus chemokine, vCXCL-1 (CCMV). Virology 364, 454-65.

Mocarski, E. S., Kemble, G. W., Lyle, J. M. and Greaves, R. F. (1996). A deletion mutant in the human cytomegalovirus gene encoding IE1(491aa) is replication defective due to a failure in autoregulation. Proc Natl Acad Sci U S A 93, 11321-6.

Mocarski, E. S., Prichard, M. N., Tan, C. S. and Brown, J. M. (1997). Reassessing the organization of the UL42-UL43 region of the human cytomegalovirus strain AD169 genome. Virology 239, 169-75.

Mocarski, E.S., and Courcelle, C.T. (2001). Cytomegaloviruses and their replication. In Fields Virology, pp. 2629-2673. Edited by P. M. H. D. M. Knipe, D. E. Griffin, R. A. Lamb, and M. A. Martin. Philadephia: Lippincott Williams and Wilkins.

Mocarski, E. S. (2006). Myeloid cell recruitment and function in pathogenesis and latency. In Cytomegaloviruses:Molecular biology and immunology, 465-481. Edited by R. MJ. Wymondham: Caister.

Mousavi-Jazi, M., Sundqvist, V. A., Linde, A., Wahren, B. and Brytting, M. (2000). Growth phenotypes of cytomegalovirus isolates do not correlate with glycoprotein B, major immediate early genotypes or antiviral sensitivity. J Med Virol 62, 117-26.

Muranyi W., Haas J., Wagner M., Krohne G. and Koszinowski U.H. (2002). Cytomegalovirus recruitment of cellular kinases to dissolve the nuclear lamina, *Science* 297 (5582), 854–857.

Murphy, J. C., Fischle, W., Verdin, E. and Sinclair, J. H. (2002). Control of cytomegalovirus lytic gene expression by histone acetylation. Embo J 21, 1112-20.

Murphy, E., Yu, D., Grimwood, J., Schmutz, J., Dickson, M., Jarvis, M. A., Hahn, G., Nelson, J. A., Myers, R. M. and Shenk, T. E. (2003). Coding potential of laboratory and clinical strains of human cytomegalovirus. Proc Natl Acad Sci U S A 100, 14976-81.

Murphy, E., Rigoutsos, I., Shibuya, T. and Shenk, T. E. (2003a). Reevaluation of human cytomegalovirus coding potential. Proc Natl Acad Sci U S A 100, 13585-90.

Nascimento, M. C., Wilder, N., Pannuti, C. S., Weiss, H. A. and Mayaud, P. (2005). Molecular characterization of Kaposi's sarcoma associated herpesvirus (KSHV) from patients with AIDS-associated Kaposi's sarcoma in Sao Paulo, Brazil. J Clin Virol 33, 52-9.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3, 418-26.

Nei, M. and Kumar, S. (2000) Molecular Evolution and Phylogenetics, Oxford University Press. Oxford.

Neote, K., DiGregorio, D., Mak, J. Y., Horuk, R. and Schall, T. J. (1993). Molecular cloning, functional expression, and signaling characteristics of a C-C chemokine receptor. Cell 72, 415-25.

Nicola, A. V., Hou, J., Major, E. O. and Straus, S. E. (2005). Herpes simplex virus type 1 enters human epidermal keratinocytes, but not neurons, via a pH-dependent endocytic pathway. J Virol 79, 7609-16.

Nishiyama, Y., Kimura, H. and Daikoku, T. (1991). Complementary lethal invasion of the central nervous system by nonneuroinvasive herpes simplex virus types 1 and 2. J Virol 65, 4520-4.

Oien, N. L., Thomsen, D. R., Wathen, M. W., Newcomb, W. W., Brown, J. C. and Homa, F. L. (1997). Assembly of herpes simplex virus capsids using the human cytomegalovirus scaffold protein: critical role of the C terminus. J Virol 71, 1281-91.

Oram, J. D., Downing, R. G., Akrigg, A., Dollery, A. A., Duggleby, C. J., Wilkinson, G. W. and Greenaway, P. J. (1982). Use of recombinant plasmids to investigate the structure of the human cytomegalovirus genome. J Gen Virol 59, 111-29.

Pass, R. F., Duliege, A. M., Boppana, S., Sekulovich, R., Percell, S., Britt, W. and Burke, R. L. (1999). A subunit cytomegalovirus vaccine based on recombinant envelope glycoprotein B and a new adjuvant. J Infect Dis 180, 970-5.

Paterson, D. A., Dyer, A. P., Milne, R. S., Sevilla-Reyes, E. and Gompels, U. A. (2002). A role for human cytomegalovirus glycoprotein O (gO) in cell fusion and a new hypervariable locus. Virology 293, 281-94.

Penfold, M. E., Dairaghi, D. J., Duke, G. M., Saederup, N., Mocarski, E. S., Kemble, G. W. and Schall, T. J. (1999). Cytomegalovirus encodes a potent alpha chemokine. Proc Natl Acad Sci U S A 96, 9839-44.

Perdue, M. L., Garcia, M., Senne, D. and Fraire, M. (1997). Virulence-associated sequence duplication at the hemagglutinin cleavage site of avian influenza viruses. Virus Res 49, 173-86.

Picone, O., Ville, Y., Costa, J. M., Rouzioux, C. and Leruez-Ville, M. (2005). Human cytomegalovirus (HCMV) short tandem repeats analysis in congenital infection. J Clin Virol 32, 254-6.

Picone, O., Costa, J. M., Chaix, M. L., Ville, Y., Rouzioux, C. and Leruez-Ville, M. (2005a). Human cytomegalovirus UL144 gene polymorphisms in congenital infections. J Clin Microbiol 43, 25-9.

Pignatelli, S., Dal Monte, P. and Landini, M. P. (2001). gpUL73 (gN) genomic variants of human cytomegalovirus isolates are clustered into four distinct genotypes. J Gen Virol 82, 2777-84.

Pignatelli, S., Dal Monte, P., Zini, N., Valmori, A., Maraldi, N. M. and Landini, M. P. (2002). Immunoelectron microscopy analysis of HCMV gpUL73 (gN) localization. Arch Virol 147, 1247-56.

Pignatelli, S., Dal Monte, P., Rossini, G., Chou, S., Gojobori, T., Hanada, K., Guo, J. J., Rawlinson, W., Britt, W., Mach, M. and Landini, M. P. (2003). Human cytomegalovirus glycoprotein N (gpUL73-gN) genomic variants: identification of a novel subgroup, geographical distribution and evidence of positive selective pressure. J Gen Virol 84, 647-55.

Pignatelli, S., Dal Monte, P., Rossini, G. and Landini, M. P. (2004). Genetic polymorphisms among human cytomegalovirus (HCMV) wild-type strains. Rev Med Virol 14, 383-410.

Pignatelli, S., Dal Monte, P., Rossini, G., Camozzi, D., Toscano, V., Conte, R. and Landini, M. P. (2006). Latency-associated human cytomegalovirus glycoprotein N genotypes in monocytes from healthy blood donors. Transfusion 46, 1754-62.

Plotkin, S. A. (2001). Vaccination against cytomegalovirus. Arch Virol Suppl, 121-34.

Poncet, C., Frances, V., Gristina, R., Scheiner, C., Pellissier, J. F. and Figarella-Branger, D. (1996). CD24, a glycosylphosphatidylinositol-anchored molecules is transiently expressed during the development of human central nervous system and is a marker of human neural cell lineage tumors. Acta Neuropathol (Berl) 91, 400-8.

Prepens, S., Kreuzer, K. A., Leendertz, F., Nitsche, A. and Ehlers, B. (2007). Discovery of herpesviruses in multi-infected primates using locked nucleic acids (LNA) and a bigenic PCR approach. Virol J 4, 84.

Prichard, M. N., Penfold, M. E., Duke, G. M., Spaete, R. R. and Kemble, G. W. (2001). A review of genetic differences between limited and extensively passaged human cytomegalovirus strains. Rev Med Virol 11, 191-200.

Pride, D. (2004). Swaap 1.0.1: a tool for analyzing substitutions and similarity in multiple alignments
[http://www.bacteriamuseum.org/SWAAP/SwaapPage.htm].

Puchhammer-Stockl, E., Gorzer, I., Zoufaly, A., Jaksch, P., Bauer, C. C., Klepetko, W. and Popow-Kraupp, T. (2006). Emergence of multiple cytomegalovirus strains in blood and lung of lung transplant recipients. Transplantation 81, 187-94.

Qi, Y., Mao, Z. Q., Ruan, Q., He, R., Ma, Y. P., Sun, Z. R., Ji, Y. H. and Huang, Y. (2006). Human cytomegalovirus (HCMV) UL139 open reading frame: Sequence variants are clustered into three major genotypes. J Med Virol 78, 517-22.

Qiu, X., Wu, L., Huang, H., McDonel, P. E., Palumbo, A. V., Tiedje, J. M. and Zhou, J. (2001). Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. Appl Environ Microbiol 67, 880-7.

Ramachandran, G. N. (1963). Protein Structure and Crystallography. Science 141, 288-291.

Rasmussen, L., Hong, C., Zipeto, D., Morris, S., Sherman, D., Chou, S., Miner, R., Drew, W. L., Wolitz, R., Dowling, A., Warford, A. and Merigan, T. C. (1997). Cytomegalovirus gB genotype distribution differs in human immunodeficiency virus-infected patients and immunocompromised allograft recipients. J Infect Dis 175, 179-84.

Rasmussen, L., Geissler, A., Cowan, C., Chase, A. and Winters, M. (2002). The genes encoding the gCIII complex of human cytomegalovirus exist in highly diverse combinations in clinical isolates. J Virol 76, 10841-8.

Rasmussen, L., Geissler, A. and Winters, M. (2003). Inter- and intragenic variations complicate the molecular epidemiology of human cytomegalovirus. J Infect Dis 187, 809-19.

Reeves, M. B., Lehner, P. J., Sissons, J. G. and Sinclair, J. H. (2005). An in vitro model for the regulation of human cytomegalovirus latency and reactivation in dendritic cells by chromatin remodelling. J Gen Virol 86, 2949-54.

Retiere, C., Imbert, B. M., David, G., Courcoux, P. and Hallet, M. M. (1998). A polymorphism in the major immediate-early gene delineates groups among cytomegalovirus clinical isolates. Virus Res 57, 43-51.

Revello, M. G. and Gerna, G. (2004). Pathogenesis and prenatal diagnosis of human cytomegalovirus infection. J Clin Virol 29, 71-83.

Rowe, W. P., Hartley, J. W., Waterman, S., Turner, H. C. and Huebner, R. J. (1956). Cytopathogenic agent resembling human salivary gland virus recovered from tissue cultures of human adenoids. Proc Soc Exp Biol Med 92, 418-24.

Ryckman, B. J., Rainish, B. L., Chase, M. C., Borton, J. A., Nelson, J. A., Jarvis, M. A. and Johnson, D. C. (2008). Characterization of the human cytomegalovirus gH/gL/UL128-131 complex that mediates entry into epithelial and endothelial cells. J Virol 82, 60-70.

Sahagun-Ruiz, A., Sierra-Honigmann, A. M., Krause, P. and Murphy, P. M. (2004). Simian cytomegalovirus encodes five rapidly evolving chemokine receptor homologues. Virus Genes 28, 71-83.

Schabath, H., Runz, S., Joumaa, S. and Altevogt, P. (2006). CD24 affects CXCR4 function in pre-B lymphocytes and breast carcinoma cells. J Cell Sci 119, 314-25.

Schaeffer, H. J., Beauchamp, L., de Miranda, P., Elion, G. B., Bauer, D. J. and Collins, P. (1978). 9-(2-hydroxyethoxymethyl) guanine activity against viruses of the herpes group. Nature 272, 583-5.

Sekulin, K., Gorzer, I., Heiss-Czedik, D. and Puchhammer-Stockl, E. (2007). Analysis of the variability of CMV strains in the RL11D domain of the RL11 multigene family. Virus Genes 35, 577-83.

Sevilla-Reyes E. (2007). Recombination in Human Cytomegalovirus. PhD thesis.

Shaw, S. B., Rasmussen, R. D., McDonough, S. H., Staprans, S. I., Vacquier, J. P. and Spector, D. H. (1985). Cell-related sequences in the DNA genome of human cytomegalovirus strain AD169. J Virol 55, 843-8.

Shepp, D. H., Match, M. E., Lipson, S. M. and Pergolizzi, R. G. (1998). A fifth human cytomegalovirus glycoprotein B genotype. Res Virol 149, 109-14.

Silva, M. C., Yu, Q. C., Enquist, L. and Shenk, T. (2003). Human cytomegalovirus UL99-encoded pp28 is required for the cytoplasmic envelopment of tegument-associated capsids. J Virol 77, 10594-605.

Sinclair, J. and Sissons, P. (2006). Latency and reactivation of human cytomegalovirus. J Gen Virol 87, 1763-79.

Sinzger, C., Hahn, G., Digel, M., Katona, R., Sampaio, K. L., Messerle, M., Hengel, H., Koszinowski, U., Brune, W. and Adler, B. (2008). Cloning and sequencing of a highly productive, endotheliotropic virus strain derived from human cytomegalovirus TB40/E. J Gen Virol 89, 359-68.

Skaletskaya, A., Bartle, L. M., Chittenden, T., McCormick, A. L., Mocarski, E. S. and Goldmacher, V. S. (2001). A cytomegalovirus-encoded inhibitor of apoptosis that suppresses caspase-8 activation. Proc Natl Acad Sci U S A 98, 7829-34.

Smith, M. G. (1956). Propagation in tissue cultures of a cytopathogenic virus from human salivary gland virus (SGV) disease. Proc Soc Exp Biol Med 92, 424-30.

Smith, J. D. & De Harven, E. (1973). Herpes simplex virus and human cytomegalovirus replication in WI-38 cells. I. Sequence of viral replication. J Virol 12, 919-30.

Smith, J. A. and Pari, G. S. (1995). Human cytomegalovirus UL102 gene. J Virol 69, 1734-40.

Smith, S. C., Oxford, G., Wu, Z., Nitz, M. D., Conaway, M., Frierson, H. F., Hampton, G. and Theodorescu, D. (2006). The metastasis-associated gene CD24 is regulated by Ral GTPase and is a mediator of cell proliferation and survival in human cancer. Cancer Res 66, 1917-22.

Soderberg-Naucler, C., Streblow, D. N., Fish, K. N., Allan-Yorke, J., Smith, P. P. and Nelson, J. A. (2001). Reactivation of latent human cytomegalovirus in CD14(+) monocytes is differentiation dependent. J Virol 75, 7543-54.

Soroceanu L., Akhavan A., and Cobbs A. (2008). Platelet-derived growth factor-α receptor activation is required for human cytomegalovirus infection. Nature, 455, 391-395.

Spaete, R. R. and Mocarski, E. S. (1985). The alpha sequence of the cytomegalovirus genome functions as a cleavage/packaging signal for herpes simplex virus defective genomes. J Virol 54, 817-24.

Spector, D. H. (1996). Activation and regulation of human cytomegalovirus early genes. Intervirology 39, 361-77.

Staden, R., Beal, K. F. and Bonfield, J. K. (2000). The Staden package, 1998. Methods Mol Biol 132, 115-30.

Stagno, S., Pass, R. F., Dworsky, M. E., Henderson, R. E., Moore, E. G., Walton, P. D. and Alford, C. A. (1982). Congenital cytomegalovirus infection: The relative importance of primary and recurrent maternal infection. N Engl J Med 306, 945-9.

Stagno, S., Dworsky, M. E., Torres, J., Mesa, T. and Hirsh, T. (1982a). Prevalence and importance of congenital cytomegalovirus infection in three different populations. J Pediatr 101, 897-900.

Stamminger, T., Gstaiger, M., Weinzierl, K., Lorz, K., Winkler, M. and 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, 4836-47.

Stanton, R., Westmoreland, D., Fox, J. D., Davison, A. J. and Wilkinson, G. W. (2005). Stability of human cytomegalovirus genotypes in persistently infected renal transplant recipients. J Med Virol 75, 42-6.

Stenberg, R. M. (1996). The human cytomegalovirus major immediate-early gene. Intervirology 39, 343-9.

Stratton, K. R. (2000). Vaccines for the 21st Century: A tool for Decision making. In National Academy Press. Washington.

Sullivan, V., Talarico, C. L., Stanat, S. C., Davis, M., Coen, D. M. and Biron, K. K. (1992). A protein kinase homologue controls phosphorylation of ganciclovir in human cytomegalovirus-infected cells. Nature 359, 85.

Suzuki, Y. and Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16, 1315-28.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Mol Biol Evol 24, 1596-9.

Tanaka, K., Numazaki, K. and Tsutsumi, H. (2005). Human cytomegalovirus genetic variability in strains isolated from Japanese children during 1983-2003. J Med Virol 76, 356-60.

Tarrago, D., Quereda, C. and Tenorio, A. (2003). Different cytomegalovirus glycoprotein B genotype distribution in serum and cerebrospinal fluid specimens determined by a novel multiplex nested PCR. J Clin Microbiol 41, 2872-7.

Taylor-Wiedeman, J., Sissons, J. G., Borysiewicz, L. K. and Sinclair, J. H. (1991). Monocytes are a major site of persistence of human cytomegalovirus in peripheral blood mononuclear cells. J Gen Virol 72 (Pt 9), 2059-64.

Taylor-Wiedeman, J., Sissons, P. and Sinclair, J. (1994). Induction of endogenous human cytomegalovirus gene expression after differentiation of monocytes from healthy carriers. J Virol 68, 1597-604.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673-80.

Tomasec, P., Braud, V. M., Rickards, C., Powell, M. B., McSharry, B. P., Gadola, S., Cerundolo, V., Borysiewicz, L. K., McMichael, A. J. and Wilkinson, G. W. (2000). Surface expression of HLA-E, an inhibitor of natural killer cells, enhanced by human cytomegalovirus gpUL40. Science 287, 1031.

Tomasec, P., Wang, E. C., Davison, A. J., Vojtesek, B., Armstrong, M., Griffin, C., McSharry, B. P., Morris, R. J., Llewellyn-Lacey, S., Rickards, C., Nomoto, A., Sinzger, C. and Wilkinson, G. W. (2005). Downregulation of natural killer cell-activating ligand CD155 by human cytomegalovirus UL141. Nat Immunol 6, 181-8.

Trincado, D. E., Scott, G. M., White, P. A., Hunt, C., Rasmussen, L. and Rawlinson, W. D. (2000). Human cytomegalovirus strains associated with congenital and perinatal infections. J Med Virol 61, 481-7.

Tu, W., Chen, S., Sharp, M., Dekker, C., Manganello, A. M., Tongson, E. C., Maecker, H. T., Holmes, T. H., Wang, Z., Kemble, G., Adler, S., Arvin, A. and

Lewis, D. B. (2004). Persistent and selective deficiency of CD4+ T cell immunity to cytomegalovirus in immunocompetent young children. J Immunol 172, 3260-7.

Urban M., Klein M., Britt W.J., Hassfurther E., Mach M. (1996). Glycoprotein H of human cytomegalovirus is a major antigen for the neutralizing humoral immune response. J Gen Virol. 77 (7), 1537-47.

Umene, K. (1999). Mechanism and application of genetic recombination in herpesviruses. Rev Med Virol 9, 171-82.

Varnum, S. M., Streblow, D. N., Monroe, M. E., Smith, P., Auberry, K. J., Pasa-Tolic, L., Wang, D., Camp, D. G., 2nd, Rodland, K., Wiley, S., Britt, W., Shenk, T., Smith, R. D. and Nelson, J. A. (2004). Identification of proteins in human cytomegalovirus (HCMV) particles: the HCMV proteome. J Virol 78, 10960-6.

Wahman, A., Melnick, S. L., Rhame, F. S. and Potter, J. D. (1991). The epidemiology of classic, African, and immunosuppressed Kaposi's sarcoma. Epidemiol Rev 13, 178-99.

Walker, A., Petheram, S. J., Ballard, L., Murph, J. R., Demmler, G. J. and Bale, J. F., Jr. (2001). Characterization of human cytomegalovirus strains by analysis of short tandem repeat polymorphisms. J Clin Microbiol 39, 2219-26.

Walter, E. A., Greenberg, P. D., Gilbert, M. J., Finch, R. J., Watanabe, K. S., Thomas, E. D. and Riddell, S. R. (1995). Reconstitution of cellular immunity against cytomegalovirus in recipients of allogeneic bone marrow by transfer of T-cell clones from the donor. N Engl J Med 333, 1038-44.

Wang, X. and Hutt-Fletcher, L. M. (1998). Epstein-Barr virus lacking glycoprotein gp42 can bind to B cells but is not able to infect. J Virol 72, 158-63.

Wang X., Huong S., Chiu M., Raab-Traub N., Huang E. (2003). Epidermal growth factor receptor is a cellular receptor for human cytomegalovirus.
Nature, 424 (6947), 456-461.

Wang, W., Patterson, C. E., Yang, S. and Zhu, H. (2004). Coupling generation of cytomegalovirus deletion mutants and amplification of viral BAC clones. J Virol Methods 121, 137-43.

Whitley, R. J., Cloud, G., Gruber, W., Storch, G. A., Demmler, G. J., Jacobs, R. F., Dankner, W., Spector, S. A., Starr, S., Pass, R. F., Stagno, S., Britt, W. J., Alford, C., Jr., Soong, S., Zhou, X. J., Sherrill, L., FitzGerald, J. M. and Sommadossi, J. P. (1997). Ganciclovir treatment of symptomatic congenital cytomegalovirus infection: results of a phase II study. National Institute of Allergy and Infectious Diseases Collaborative Antiviral Study Group. J Infect Dis 175, 1080-6.

Wilkinson GW, Tomasec P, Stanton RJ, Armstrong M, Prod'homme V, Aicheler R, McSharry BP, Rickards CR, Cochrane D, Llewellyn-Lacey S, Wang EC, Griffin CA, Davison AJ. (2008). Modulation of natural killer cells by human cytomegalovirus. J Clin Virol. 41(3), 206-12.

Wood LJ, Baxter MK, Plafker SM, Gibson W. (1997). Human cytomegalovirus capsid assembly protein precursor (pUL80.5) interacts with itself and with the major capsid protein (pUL86) through two different domains. J Virol. 71(1), 179-90.

Xiong, X., Smith, J. L. and Chen, M. S. (1997). Effect of incorporation of cidofovir into DNA by human cytomegalovirus DNA polymerase on DNA elongation. Antimicrob Agents Chemother 41, 594-9.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A. M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431-49.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19, 908-17.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24, 1586-91.

Yu, D., Smith, G. A., Enquist, L. W. and 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, 2316-28.

Zaia, J. A. (2002). Prevention of cytomegalovirus disease in hematopoietic stem cell transplantation. Clin Infect Dis 35, 999-1004.

Zhang, H., Thorgaard, G. H. and Ristow, S. S. (2002). Molecular cloning and genomic structure of an interleukin-8 receptor-like gene from homozygous clones of rainbow trout (Oncorhynchus mykiss). Fish Shellfish Immunol 13, 251-8.

Zipeto, D., Hong, C., Gerna, G., Zavattoni, M., Katzenstein, D., Merigan, T. C. and Rasmussen, L. (1998). Geographic and demographic differences in the frequency of human cytomegalovirus gB genotypes 1-4 in immunocompromised patients. AIDS Res Hum Retroviruses 14, 533-6.