

Genotypic Analysis of Two Hypervariable Human Cytomegalovirus Genes

Amanda J. Bradley,¹ Ida J. Kovács,² Derek Gatherer,¹ Derrick J. Dargan,¹ Khaled R. Alkharsah,³ Paul K.S. Chan,⁴ William F. Carman,⁵ Martin Dedicoat,⁶ Vincent C. Emery,⁷ Colin C. Geddes,⁸ Giuseppe Gerna,⁹ Bassam Ben-Ismaeil,⁵ Steve Kaye,¹⁰ Alistair McGregor,¹¹ Paul A. Moss,¹² Rozalia Pusztai,² William D. Rawlinson,¹³ Gillian M. Scott,¹³ Gavin W.G. Wilkinson,¹⁴ Thomas F. Schulz,³ and Andrew J. Davison^{1*}

¹MRC Virology Unit, Institute of Virology, University of Glasgow, Glasgow, United Kingdom

²Department of Medical Microbiology and Immunobiology, University of Szeged, Szeged, Hungary

³Institut für Virologie, Medizinische Hochschule Hannover, Hannover, Germany

⁴Department of Microbiology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong Special Administrative Region, China

⁵West of Scotland Specialist Virology Centre, Gartnavel General Hospital, Glasgow, United Kingdom

⁶Ngwelezane Hospital, Empangeni, KwaZulu/Natal, South Africa

⁷Centre for Virology, Division of Infection and Immunity, Royal Free and University College Medical School, Hampstead, London, UK

⁸Renal Unit, Western Infirmary, Glasgow, United Kingdom

⁹Servizio di Virologia, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

¹⁰Viral Disease Programme, Medical Research Council Laboratories, Atlantic Boulevard, Banjul, The Gambia

¹¹Division of Infectious Diseases, University of Minnesota Department of Pediatrics, Center for Infectious Diseases and Microbiology Translational Research, McGuire Translational Research Facility, Minneapolis, Minnesota

¹²Cancer Research UK Institute for Cancer Studies, The University of Birmingham, Edgbaston, Birmingham, United Kingdom

¹³Virology Division, Department of Microbiology, Prince of Wales Hospital, Randwick, New South Wales, Australia

¹⁴Department of Medical Microbiology, Wales School of Medicine, Cardiff University, Heath Park, Cardiff, United Kingdom

Most human cytomegalovirus (HCMV) genes are highly conserved in sequence among strains, but some exhibit a substantial degree of variation. Two of these genes are UL146, which encodes a CXC chemokine, and UL139, which is predicted to encode a membrane glycoprotein. The sequences of these genes were determined from a collection of 184 HCMV samples obtained from Africa, Australia, Asia, Europe, and North America. UL146 is hypervariable throughout, whereas variation in UL139 is concentrated in a sequence encoding a potentially highly glycosylated region. The UL146 sequences fell into 14 genotypes, as did all previously reported sequences. The UL139 sequences grouped into 8 genotypes, and all previously reported sequences fell into a subset of these. There were minor differences among continents in genotypic frequencies for UL146 and UL139, but no clear geographical separation, and identical nucleotide sequences were represented among communities distant from each other. The frequent detection of multiple genotypes indicated that mixed infections are common. For both genes, the degree of divergence was sufficient to preclude reliable sequence alignments between genotypes in the most variable regions, and the mode of evolution

involved in generating the genotypes could not be discerned. Within genotypes, constraint appears to have been the predominant mode, and positive selection was detected marginally at best. No evidence was found for linkage disequilibrium. The emerging scenario is that the HCMV genotypes developed in early human populations (or even earlier), becoming established via founder or bottleneck effects, and have spread, recombined and mixed worldwide in more recent times. **J. Med. Virol.** 80:1615–1623, 2008. © 2008 Wiley-Liss, Inc.

KEY WORDS: herpesvirus; variation; genotype; chemokine; glycoprotein

Grant sponsor: UK Medical Research Council (to A.J.B., D.G., D.J.D., S.K., G.W.G.W., A.J.D.); Grant sponsor: Wellcome Trust (to M.D., V.C.E., G.W.G.W.); Grant sponsor: Fondazione CAR-IPLO, Milan, Italy (to G.G.).

*Correspondence to: Andrew J. Davison, MRC Virology Unit, Institute of Virology, University of Glasgow, Church Street, Glasgow G115JR, UK. E-mail: a.davison@mrcvu.gla.ac.uk

Accepted 6 May 2008

DOI 10.1002/jmv.21241

Published online in Wiley InterScience
(www.interscience.wiley.com)

INTRODUCTION

Human cytomegalovirus (HCMV; family *Herpesviridae*, subfamily *Betaherpesvirinae*, genus *Cytomegalovirus*, species *Human herpesvirus 5*) is ubiquitous and host-specific. Infection is asymptomatic for most people, but can result in serious disease in immunocompromised individuals and congenitally infected newborns. The minimally passaged Merlin strain is considered best to represent wild-type HCMV, and has a 236 kbp genome that is predicted to contain approximately 165 protein-encoding genes [Dolan et al., 2004].

Most genes are highly conserved in sequence between HCMV strains, but a number of genes predicted to encode membrane-associated or secreted proteins are characterized by a striking degree of variability, as revealed by examination of individual genes [reviewed in Pignatelli et al., 2004] and by whole genome comparisons [Murphy et al., 2003; Dolan et al., 2004]. Various studies have attempted to make connections between the genotypes of hypervariable genes and disease outcome, and overall the conclusions reached are unclear or contradictory [reviewed in Puchhammer-Stöckl and Görzer, 2006]. The apparently random association of genotypes at different loci (that is, the absence of linkage disequilibrium), presumably as a result of recombination during HCMV evolution, limits any conclusions to the specific gene under study, except in some cases where genes are very close to each other or encode interacting hypervariable proteins [Rasmussen et al., 2002, 2003]. A further complicating factor is the occurrence of multiple HCMV genotypes in individuals, including immunocompromised patients, pregnant women and congenitally infected newborns [for recent papers, see Rasmussen et al., 2003; Hassan-Walker et al., 2004; Stanton et al., 2005; Puchhammer-Stöckl et al., 2006; Iwasenko et al., 2007].

One of the most variable HCMV genes is UL146, which encodes a chemokine designated vCXC-1. This gene is variable throughout its length [Penfold et al., 1999; Prichard et al., 2001; Hassan-Walker et al., 2004; Arav-Boger et al., 2005, 2006; Stanton et al., 2005; He et al., 2006; Lurain et al., 2006], and 14 genotypes have been catalogued [Dolan et al., 2004]. In strain Toledo, UL146 encodes a functional chemokine that is capable of neutrophil degranulation, chemotaxis and calcium mobilization. This protein contains an ELRCXC motif, which has been shown to be essential for receptor binding and IL-8 activity [Clark-Lewis et al., 1991]. vCXC-1 binds to human CXCR2 and is comparable in its activities to CXC chemokines IL-8 and gro- α [Penfold et al., 1999]. The function of vCXC-1 may be to facilitate dissemination of the virus through its ability to attract monocytes to the initial site of infection. Thus, the virus could undermine the effectiveness of antiviral immunity by manipulating the host chemokine system and, together with other virus-encoded molecular mimics, suppressing the immune system.

The most variable gene in the vicinity of UL146 is UL139, which is located 5.2 kbp distant and is predicted

to encode a type I membrane glycoprotein. Variability is concentrated in a region of the ectodomain [Dolan et al., 2004]. A recent study of 26 HCMV strains isolated in China described three major genotypes, with two of these divided into subtypes [Qi et al., 2006].

The aims of the present study were to investigate whether additional UL146 genotypes exist in a large number of clinical samples obtained from a wide range of locations and clinical settings, and to define the range of UL139 genotypes in these samples. Ancillary interests were to examine the relative frequencies and geographical distribution of genotypes, to assess whether infections with more than one HCMV strain are common, and to investigate the evolution of UL146 and UL139.

MATERIALS AND METHODS

Virus DNA Collection

A collection of 184 DNA samples was derived from 179 anonymized clinical samples obtained in various geographical locations in accordance with local ethical guidelines, plus 5 commonly used laboratory strains (Davis, Merlin, TB40/E, Toledo and Towne). Details of the 171 samples in the collection that yielded sequence data are available on request, and include the age, sex, and pathology of the patient, the clinical source of the sample, and the UL146 and UL139 genotypes determined. The samples numbered 18 from Australia, 10 from Hong Kong, 6 from Germany, 13 from England, 18 from The Gambia, 24 from Hungary, 7 from Italy, 6 from The Netherlands, 41 from Scotland, 5 from the USA, 8 from Wales, and 15 from South Africa. A minority of strains (40) had been passaged in human fibroblast cell culture, either as routine diagnostic specimens or as laboratory strains. DNA was extracted by standard methods from body tissues, urine, saliva or infected cells. The South African samples were obtained from the saliva of mothers (10 of whom tested HIV-negative) attending rural clinics in KwaZulu/Natal [Dedicoat et al., 2004]. Since these were available in very limited amounts and contained low numbers of HCMV genomes, whole genome amplification using a REPLI-g kit (Qiagen, Crawley, UK) was carried out prior to PCR amplification.

PCR Amplification

UL146 and UL139 were amplified separately by single round or nested PCR, using primers in conserved regions (Table I). 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. UL140-11A is located within the UL139 coding region, and as a consequence the sequences obtained using nested PCR (approximately 40% of the total) lacked 29

TABLE I. Primers Used for PCR and Sequencing

Gene	Primer	Sequence (5'–3')	Genome location ^a
UL146	AB4	TAGACACTACGTCGTAAATG	180494–180513
UL146	A162	TGTAGAATTAGTCTAGATTCCTGA	181524–181501
UL146	UL146-4A	GCTTGC GCGTTAGGATTGAGACAC	180571–180594
UL146	UL146-3A	ATACCGGATATTACGAATT	181341–181323
UL139	AB1	GTCATTGTGAAAGTGACGTCTCAG	186389–186412
UL139	AB2	ATCTACTGTAAACCCTCTGCTCTG	187148–187125
UL139	UL140-11A	GCGGCATTGGTGTACGCGTG	186553–186572
UL139	UL140-3A	GTGGAAATTTTACGTCATT	187077–187058

^aWith reference to RefSeq accession NC_006273.2 (HCMV strain Merlin).

amino acid-encoding codons from the highly conserved C terminus.

For the single (and first) round, 1 µl of DNA was added to the PCR reaction mixture, which consisted of 40 µl of water, 5 µl of buffer, 1 µl of 10 µM dNTPs, 1 µl of each the two primers (10 µM) and 1 µl (1 U) of DNA polymerase (Advantage 2, BD Clontech, Basingstoke, UK). The conditions for amplification were 95°C for 2 min followed by 35 cycles of 95°C for 2 min, 60°C for 30 sec and 68°C for 1 min. 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 on three separate occasions to assess reproducibility.

Purification, Cloning, and Sequencing of PCR Products

PCR products were separated by agarose gel electrophoresis. Appropriate DNA fragments were excised, purified using a GeneClean turbo kit (Q Biogene, Cambridge, UK), and eluted using 100 µl of nuclease-free water. The single round or second round primers were used for direct sequencing.

In some cases, including those where direct sequencing indicated the presence of more than one genotype of UL146 or UL139, fragments were cloned using a pGEM-T kit (Promega, Southampton, UK). Following ligation and transformation into chemically competent *E. coli* TOP 10 cells, 5 recombinant colonies were picked and grown overnight at 37°C in 2YT-broth containing 100 µg/ml ampicillin. Plasmid DNA was purified using a QIAprep Spin miniprep kit (Qiagen). Plasmid inserts were sequenced using universal forward and reverse primers. Sequencing was carried out on both DNA strands using a BigDye terminator kit (Applied Biosystems, Warrington, UK) in an ABI 3730 instrument. Samples containing multiple strains were identified by the derivation of plasmids representing different genotypes of UL146 or UL139.

Sequence Analysis

Sequence chromatograms were viewed using Editview (Applied Biosystems) and analyzed using Pregap4 and Gap4 [Staden et al., 2000]. Nucleotide and imputed

amino acid sequences were aligned using CLUSTAL W [Thompson et al., 1994] and MAFFT [Katoh et al., 2005]. Full-length sequences were used for the UL146 data and a subset of the UL139 data, and another subset of the UL139 data was analyzed using sequences lacking the conserved C terminus. MEGA4.0 [Tamura et al., 2007] was used for the generation of phylogenetic trees. Frequencies of nonsynonymous and synonymous differences per site (dN and dS, respectively) and degree of sequence variability (nucleotide and amino acid) were investigated using Swaap 1.0.1 [Pride, 2004], DnaSP 4.10.9 [Rozas et al., 2003], and MEGA4.0. dN/dS ratios and probabilities of positive selection were assessed using PAML 3.15 [Yang, 1997]. Signal peptide and transmembrane sequences were predicted using Phobius [Kall et al., 2004].

Statistical Analysis

Sample origin was divided into four regions (Africa, Asia, Europe, and Australia) for assessment of the geographical distribution of genotypes. Chi-square tests were used to assess the significance of variability of genotype frequencies among regions. Yates' correction for continuity was applied to chi-square tests in cases where the expected values fell below 5. Similarly, Chi-square tests with Yates' correction were applied to 60 samples where single genotypes were detected for both UL146 and UL139, in order to test for linkage disequilibrium. Samples containing mixed infections were excluded from this analysis.

RESULTS

UL146 and UL139 Sequences

The UL146 and UL139 genotypes in 184 samples were investigated by PCR and sequencing using primers in conserved regions. UL146 was amplified from 159 samples and sequences were determined from 134, and UL139 was amplified from 168 samples and sequences determined from 131. A total of 13 samples failed to yield products from either gene. Since some samples contained more than one virus strain, totals of 182 UL146 sequences and 183 UL139 sequences were obtained. Alignment and phylogenetic analyses involved the 350 UL146 sequences and 300 UL139 sequences derived

from the present study or reported by others in the literature [Cha et al., 1996; Davison et al., 2003; Dolan et al., 2004; Arav-Boger et al., 2005, 2006; Stanton et al., 2005; He et al., 2006; Lurain et al., 2006; Qi et al., 2006] or public sequence databases [AY999242-AY999271, AY805250-AY805303, AY818250-AY818255, Mao et al.; DQ229942-DQ229948, Ruan et al.; DQ180366, DQ180358, DQ180374, DQ180386, Zhou et al.].

UL146 Genotypes

The UL146 coding sequences range in length from 114 to 126 codons, and phylogenetic analyses indicated that all fall into the 14 genotypes defined previously and designated G1–G14 [Dolan et al., 2004]. Amino acid sequence variation among genotypes is high ($p = 0.521$, where p is protein diversity from MEGA4.0), whereas within each genotype it is low ($p = 0-0.051$ with a mean of 0.017) (Table II).

Differences in overall genotypic frequencies were observed (Table II). For example, G7 was detected in 16% of samples and G6 in fewer than 1%.

UL139 Genotypes

The UL139 coding sequences range in length from 124 to 148 codons, and phylogenetic analyses indicated that all fall into 8 genotypes designated G1–G8. Figure 1A shows a predicted amino acid sequence alignment of the primary translation products of one representative of each genotype. Figure 1B shows a phylogenetic tree constructed using these sequences.

The protein encoded by each HCMV UL139 genotype contains a putative signal peptide sequence and a transmembrane region. Variation is concentrated in the N-terminal portion of the protein. Amino acid sequence variation between genotypes is high

($p = 0.275$), whereas within each genotype it is low ($p = 0.007-0.095$ with a mean of 0.025) (Table III). Variation within genotypes tends to be higher in UL139 than in UL146, but that among genotypes is lower. Sequences in G1 exhibit a greater level of variation than those in the other genotypes.

Differences in overall genotypic frequencies were observed (Table III). For example, G2 was detected in 27% of samples and G8 in fewer than 3%.

Assessment of Positive Selection

In order to assess positive selection (i.e., for amino acid sequence diversity), the dN/dS ratio was calculated for each UL146 and UL139 genotype (Tables II and III). Positive selection was detected at the 1% significance level only in UL139 G1, and at the 5% level in UL146 G7 and G1. UL139 G6 and G7 and UL146 G2 and G13 also had values of dN/dS > 1, but these were not statistically significant. Only in UL139 G1 was an amino acid residue identified as under positive selection, although this was at position 12 in the predicted signal peptide sequence. Thus, evidence for positive selection is marginal, and it seems unlikely that this mode of diversification has featured in the evolution of UL139 and UL146 since the genotypes arose. No strong evidence emerged for positive selection in formal comparisons among genotypes (i.e., as a factor in emergence of the genotypes), but it must be registered that variation was so large as to confound reliable sequence alignments.

Geographical Distribution of Genotypes

The sequence data derived in the present work were divided into four groups representing strains obtained from Africa, Asia, Australia, and Europe. Insufficient sample numbers were obtained from America to war-

TABLE II. UL146 Diversity

Genotype	Samples	Frequency (%)	Alignment length ^a		Diversity		dN/dS ^d
			DNA	Protein	DNA ^b	Protein ^c	
G1	34	9.71	345	115	0.011	0.026	1.19 ^e
G2	25	7.14	360	120	0.002	0.005	1.48
G3	10	2.86	375	125	0.010	0.016	0.50
G4	8	2.29	369	123	0.004	0.006	0.26
G5	16	4.57	348	116	0.007	0.012	0.71
G6	2	0.57	351	117	0.029	0.051	ND
G7	57	16.3	354	118	0.011	0.015	1.29 ^e
G8	22	6.29	342	114	0.006	0.008	0.38
G9	49	14	351	117	0.017	0.032	0.94
G10	12	3.43	291	97	0.003	0.005	0.30
G11	19	5.43	339	113	0.007	0.015	0.50
G12	43	12.3	354	118	0.016	0.018	0.45
G13	47	13.4	357	119	0.007	0.015	1.58
G14	6	1.71	354	118	0	0	ND
All	350	100	225	75	0.642	0.521	0.27 ^f

^aGaps removed.

^bJukes-Cantor P_i from DnaSP 4.10.9.

^cProtein diversity p from MEGA4.0.

^ddN/dS (omega) from PAML 3.15 under the single-rate model. ND, not determined.

^eFive percent significance for positive selection.

^fCalculated from a comparison of a single member of each genotype.

A

G1	(w9)	MLWILVLFALAASA <u>SETTTGTSSNSSQS</u> TSAGTTNTTTPS . . . TACINASNGSDLGAPQLALLAAS
G2	(E8)	MLWILVLFALAASA <u>SETTTGTSSNSSQSTSSSSSSSTSS</u> . . NSTATPTSASIQCVESFGGSNWTVAQLALFAAS
G3	(E11)	MLWILVLFALATS <u>SETTTGTSSNSSQSSTSSSSSTNTSNNTTSATTLS</u> . . . TECINGFGGNNWTFPQLALFAAS
G4	(U4)	MLWILVLFALAASA <u>SETTTGTSSNSSQS</u> TSA.TANTTVS TCINASNGSSWTFPQLALLAAS
G5	(U5)	MTVVVMLTI AVAAVAIVS . <u>SNNNTTNS</u> TTCVDGTNGTWTFVQHVGMLAAG
G6	(W8)	MLWILALLALTATA <u>SETTTGTSSNSSTSTM</u> SSNSTVAPTTPS . . . VACVQAFGGSNWTFPQLALLAAS
G7	(E12)	MLWILVLFALAASA <u>SETTTGTSSNSSQATSSSSSSSTSSSNNSTATPT</u> . . . IECVQAFGGSNWTVAQLALFAAS
G8	(A3)	MLWILVLFALAASA <u>SETTTGTSSNSSQS</u> TSVTTSS . . . TACINGSGGSNWTFPQLALLAAS
con		-----A-----N-----C-----G-----AA-
CCMV		MTVTVTLVALSSAVSAAL <u>SETTTGTSSNSSQSTSS</u> TATTGT . GCSNANDTNNNGLNNQQQIIAGLLG

G1	(w9)	GWTLGGLLLIFTCCLCCFWLVRKVCSS . CCGNSSESESK . . ATHAYTNAAFTSSDATLPMGTTGSGYTPP . . QDGSFPPPPR
G2	(E8)	GWTLGGLLLFTCCFCCFWLVRKICSS . CCGNSSESESK . . TTHAYTNAAFTSSDATLPMGTTGSGYTPP . . QDGSFPPPPR
G3	(E11)	GWTLGGLLLFTCCFCCFWLVRKICSS . CCGNSSESESK . . TTHAYTNAAFTSSDATLPMGTTGSGYTPP . . QDGSFPPPPR
G4	(U4)	GWTLGGLLLFTCCFCCFWLVRKICSS . CCGNSSESESK . . TTHAYTNAAFTSSDATLPMGTTGSGYTPP . . QDGSFPPPPR
G5	(U5)	GWSCPIILLMFVCCFCCFQLLRKLCG . CCGNS . QSDSK . . TTHAYTNAAFTSSDATLPMGTTGSGYTPP . . QDGSFPPPPR
G6	(W8)	GWTLGGLLLFTCCFCCFWLVRKICSS . CCGNSSESESK . . TTHAYTNAAFTSSDATLPMGTTGSGYTPP . . QDGSFPPPPR
G7	(E12)	GWTLGGLLLFTCCFCCFWLVRKICSS . CCGNSSESESK . . TTHAYTNAAFTSSDATLPMGTTGSGYTPP . . QDGSFPPPPR
G8	(A3)	GWTLGGLLLFTCCFCCFWLVRKICSS . CCGNSSESESK . . TTHAYTNAAFTSSDSTLPMGTTGSKTTPP . . QDGSFPPPPA
con		GW - - - LLL - F - CC - CCF - L - RK - C - - CCGNS - - S - SK - - - THAYTNAAFTSSD - TLPMGTTGS - TTPP . . QDGSFPPPP -
CCMV		GCGLSLFFIIFTCILCVWVYCFRKLFPDCCGGDPDQQRQMRGRYTYDNPVFPFPTLPMGATGPAYPPPVSDGTAGPPAI

B

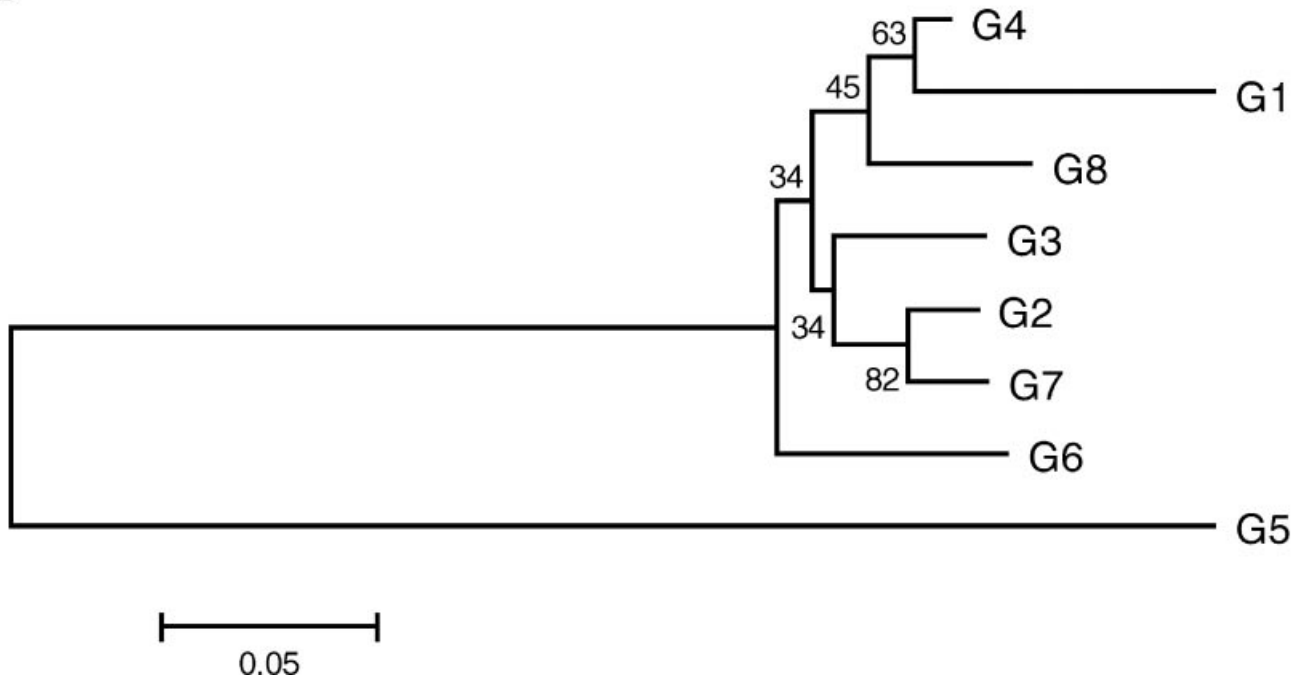


Fig. 1. Phylogenetic analysis of UL139. **A:** Alignment (CLUSTAL W) of amino acid sequences representing the eight genotypes. Predicted signal peptide and transmembrane sequences are highlighted in gray. Completely conserved residues are indicated in the consensus row (con). Below this is the CCMV sequence, which is included to illustrate conservation of the SETTTGTSSNSS motif (underlined). The CCMV sequence [Davison et al., 2003] provided is the C-terminal portion (final 12 residues not shown) of a larger protein, the N-terminal portion of

which lacks a counterpart in HCMV but is related to a protein (encoded by gene rh174) in rhesus cytomegalovirus. **B:** Unrooted neighbor-joining tree for the HCMV amino acid sequences shown in (A), computed using Mega4.0 (Poisson correction method with gaps removed). Bootstrap values (out of 100) are shown, and values below 70 indicate regions of unresolved branching order. The scale bar indicates the number of amino acid substitutions per site.

rant inclusion. Observation of frequencies initially suggested no significant differences in the distribution of UL146 and UL139 genotypes among continents (Tables IV and V). However, this conclusion may have

been compromised by insufficient sample numbers from some areas (particularly Asia) and lack of information on the ethnic origin of samples. This shortcoming is highlighted by the observation that regions (specifically

TABLE III. UL139 Diversity

Genotype	Samples	Frequency (%)	Alignment length ^a		Diversity		dN/dS ^d
			DNA	Protein	DNA ^b	Protein ^c	
G1	48	16	255	85	0.059	0.095	0.82 ^e
G2	82	27.33	240	80	0.009	0.009	0.76
G3	29	9.66	339	113	0.014	0.013	0.34
G4	68	22.66	201	67	0.023	0.015	0.76
G5	28	9.33	255	124	0.018	0.024	0.65
G6	24	8	312	104	0.010	0.013	1.08
G7	14	4.66	237	79	0.006	0.015	2.38
G8	7	2.33	228	140	0.007	0.007	0.19
All	300	100	153	51	0.285	0.275	0.48 ^f

^aGaps removed.

^bJukes-Cantor P_i from DnaSP 4.10.9.

^cProtein diversity p from MEGA4.0.

^ddN/dS (omega) from PAML 3.15 under the single-rate model.

^eOne percent significance for positive selection.

^fCalculated from a comparison of a single member of each genotype.

Europe) for which a larger sample size was tested displayed greater genotypic diversity. Nonetheless, UL146 G13 appears somewhat more common in African samples than in European samples (11 out of 45 sequences were detected in the former and 17 out of 104 in the latter), UL146 G10 and G11 appear to be restricted to Europe (8 and 3 samples, respectively), and the single sample of UL146 G6 originated from Asia.

Identical nucleotide sequences were frequently obtained from geographically distant and presumably epidemiologically unrelated patients. For example, certain samples from The Gambia, Scotland, and Hungary contained identical UL146 G12 sequences. Also, UL139 G2, which was identified in 27% of samples, was represented by identical sequences from Hungary, The UK, and The Gambia.

Linkage Disequilibrium Between UL146 and UL139

Potential linkage disequilibrium was investigated in 60 strains for which single genotypes of both UL146 and

UL139 were obtained. Of 112 possible genotype pairs, 41 were observed at least once (Table VI). No statistical significance was obtained for the observed distribution of genotype pairs versus a null hypothesis of independent assortment, indicating an absence of linkage disequilibrium.

Infections With Multiple Strains

Multiple genotypes in one or both genes were detected in at least 14% of samples upon first analysis (rising to 29% when repeat experiments were included), distributed among immunocompetent and immunocompromised individuals. 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 (rising to 24%, 33%, 60%, and 60%, respectively, when repeat experiments are included).

DISCUSSION

This study focused on the genotype definitions, frequencies, occurrence in mixed infections, geographical distribution and evolution (in terms of linkage disequilibrium and mode of selection) of two hyper-variable HCMV genes, UL146 and UL139. Totals of 182 UL146 and 183 UL139 sequences were obtained from a

TABLE IV. 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 = 177	45	13	104	15

TABLE V. 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 = 178	41	18	100	19

TABLE VI. Analysis of Linkage Disequilibrium

UL146 genotype	UL139 genotype							
	G1	G2	G3	G4	G5	G6	G7	G8
G1	1	0	0	3	1	0	0	0
G2	1	1	0	0	0	0	0	0
G3	1	0	0	0	0	0	0	0
G4	0	0	1	1	0	0	0	1
G5	0	1	0	1	0	0	1	0
G6	0	0	0	0	0	0	0	0
G7	1	4	1	3	2	1	0	0
G8	0	1	0	1	1	0	1	0
G9	0	1	0	0	0	2	1	0
G10	0	3	0	1	0	0	0	0
G11	0	2	0	1	0	0	0	0
G12	1	1	1	1	2	1	0	0
G13	2	2	0	5	1	0	0	1
G14	1	0	0	0	0	0	0	0
Totals = 60	8	16	3	17	7	4	3	2

large panel of clinical isolates collected from Africa (South Africa and The Gambia), Asia (Hong Kong), Australia and Europe (various countries). These were used in all analyses, and were supplemented by 168 previously published UL146 and 117 UL139 sequences in analyses of genotype definitions, frequencies and mode of selection.

The UL146 sequences fell into the 14 genotypes described previously [Dolan et al., 2004], and no new genotypes were discovered. Twelve genotypes contained the ELRCXC motif, which has been shown to be essential for receptor binding and IL-8 activity [Clark-Lewis et al., 1991], and 2 contained the NGRXC motif, which has been shown to be important for interaction with T and B cells [Baggiolini et al., 1997]. The latter genotypes (G5 and G6) are relatively rare, being present in approximately 5% of samples. It is not known whether the UL146 genotypes possess different biological properties, and studies to investigate this question are required.

The 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. Thus, apart from the differences in nomenclature, the subgroups [Qi et al., 2006] correlate with a subset of the genotypes in the present study, except that the closely related subgroups G1b and G1c in the former are amalgamated as G1 in the latter. Most of the variation in UL139 is due to substitutions or deletions of variable size near the N terminus. This region is rich in S and T residues that are potentially susceptible to *O*-glycosylation, and also contains NXS or NXT motifs that are potentially susceptible to *N*-glycosylation. This

suggests that selection may have focused primarily on glycosyl side chains rather than the underlying amino acid sequence. A similar feature characterizes other variable glycoprotein genes, such as UL73 (encoding glycoprotein N (gN)) and UL74 (encoding glycoprotein O (gO)) [Pignatelli et al., 2001, 2003; Mattick et al., 2004].

A region of sequence identity (SETTTGTSSNSS in Fig. 1A) has been noted between the HCMV UL139 protein and CD24, a cellular glycosyl phosphatidylinositol-linked glycoprotein that is involved in B cell activation [Qi et al., 2006]. It is difficult to assess the significance of this observation, especially since 9 of the 12 residues are S or T and the region is not conserved in CD24 orthologues from other mammals. However, the sequence is present in all of the UL139 genotypes identified in the present study, except for G5, and also in the homologous protein from chimpanzee cytomegalovirus (CCMV) (Fig. 1A). Variation in glycosylation has been observed in CD24 and has been linked to differences in cell and tissue specificity [Poncet et al., 1996]. 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 [Schabath et al., 2006; Smith et al., 2006]. The possibility that UL139 may be a CD24 homologue remains intriguing, but, in the absence of functional data, unproven.

Studies of HCMV genotype frequency, including the present one, are usually based on the use of conserved PCR primers, and face limitations as a result. Firstly, there is no guarantee that all genotypes will be detected, since primers are chosen on the basis of alignments of available sequences. Secondly, samples containing more than one strain yield mixed sequences, which when cloned are recovered approximately in proportion to their abundance (although stochastic processes may introduce bias during PCR). Therefore, the absence of a genotype from a particular sample cannot be assured. If any UL146 or UL139 genotypes have escaped recognition, they may emerge from future studies involving different primers or from whole genome sequencing exercises.

As found in previous studies [reviewed in Puchhammer-Stöckl and Görzer, 2006], mixed infections with different HCMV strains were common. In some samples, a single UL139 genotype and multiple UL146 genotypes, or vice versa, were detected. 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 sequences present as mixtures in unequal proportions. Some samples tested more than once were found to contain additional genotypes not detected in the first experiment, suggesting that the number of mixed infections was underestimated by the methodology used. Mixed infections were more frequently detected 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 of HCMV in children in Malawi compared to European countries and the USA was taken as

possibly reflecting greater opportunities for transmission, although multiple genotypes were detected in only a small number of samples.

The occurrence of mixed infections is being recognized increasingly as potentially significant to the biology of HCMV. This feature adds to the limitations inherent in studies of whether particular genotypes are associated with disease outcome; other features include the number, origin and pathological categorization of samples, the choice of gene, the absence of linkage disequilibrium, and host factors. In light of these limitations, our opinion is that robust evidence in favor of any association between genotype and pathology has proved elusive in the literature. Further work utilizing genotype-specific approaches is required to explore the true frequency of mixed infections, both to validate studies of this type and to determine whether mixed infections have geographical or biological correlates.

Similar to the conclusions drawn from a study on UL73 (encoding gN) [Pignatelli et al., 2003], no statistically significant association of UL146 or UL139 genotypes with geographical origin arose from the analysis. However, this may reflect low sample numbers (albeit much larger than those utilized in previous studies on UL146 and UL139) and the lack of detailed information on ethnic origin. Likewise, investigation of linkage disequilibrium between UL146 and UL139 genotypes was compromised by the small sample number (60) relative to the large number of possible genotype combinations (112). However, no evidence for linkage disequilibrium was obtained, indicating the involvement of recombination in HCMV evolution since the genotypes arose. Taking into account the size of the HCMV gene complement, we agree with the view that very many strains are likely to be circulating in the world [Rasmussen et al., 2003].

The extensive divergence between genotypes and the consequent inability to produce reliable sequence alignments for both UL146 and UL139 in the hypervariable regions compromised assessments of the role of positive selection in generating the genotypes. In contrast, variation within genotypes is low, and identical nucleotide sequences were obtained from geographically distant individuals. The analysis suggests that constraint has been the predominant factor in evolution within genotypes, with positive selection detected only marginally. A previous study [Arav-Boger et al., 2005] involving 30 sequences also concluded that UL146 has evolved under constraint. The sequences of hypervariable genes are stable on short timescales in patients [Hassan-Walker et al., 2004; Stanton et al., 2005] and cell culture [Lurain et al., 2006], consistent with the perception of herpesvirus genomes as relatively slowly evolving [McGeoch et al., 2006]. The most likely scenario for the evolution of HCMV emerging from the literature and from the present study is that the genotypes developed in early human populations (or even earlier), becoming established via founder or bottleneck effects, and have spread, recombined and mixed worldwide in more recent times, with mixed infections being common.

ACKNOWLEDGMENTS

IJK was a recipient of a FEMS Research Fellowship and a FEMS-ESCMID Joint Fellowship, and KRA was a recipient of a DAAD Fellowship (German Academic Exchange Service). We thank Mark Schleiss for providing the virus from which one of the samples (a BAC) was generated. We are grateful to Duncan McGeoch for comments on a draft of the manuscript.

REFERENCES

- Arav-Boger R, Foster CB, Zong JC, Pass RF. 2006. Human cytomegalovirus-encoded α -chemokines exhibit high sequence variability in congenitally infected newborns. *J Infect Dis* 193:788–791.
- Arav-Boger R, Zong JC, Foster CB. 2005. Loss of linkage disequilibrium and accelerated protein divergence in duplicated cytomegalovirus chemokine genes. *Virus Genes* 31:65–72.
- Baggiolini M, Dewald B, Moser B. 1997. Human chemokines: An update. *Annu Rev Immunol* 15:675–705.
- Beyari MM, Hodgson TA, Kondowe W, Molyneux EM, Scully C, Porter SR, Teo CG. 2005. Inter- and intra-person cytomegalovirus infection in Malawian families. *J Med Virol* 75:575–582.
- Cha TA, Tom E, Kemble GW, Duke GM, Mocarski ES, Spaete RR. 1996. Human cytomegalovirus clinical isolates carry at least 19 genes not found in laboratory strains. *J Virol* 70:78–83.
- Clark-Lewis I, Schumacher C, Baggiolini M, 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* 266:23128–23134.
- Davison AJ, Dolan A, Akter P, Addison C, Dargan DJ, Alcendor DJ, McGeoch DJ, Hayward GS. 2003. The human cytomegalovirus genome revisited: Comparison with the chimpanzee cytomegalovirus genome. *J Gen Virol* 84:17–28.
- Dedicoat M, Newton R, Alkharsah KR, Sheldon J, Szabados I, Ndlovu B, Page T, Casabonne D, Gilks CF, Cassol SA, Whitby D, Schulz TF. 2004. Mother-to-child transmission of human herpesvirus-8 in South Africa. *J Infect Dis* 190:1068–1075.
- Dolan A, Cunningham C, Hector RD, Hassan-Walker AF, Lee L, Addison C, Dargan DJ, McGeoch DJ, Gatherer D, Emery VC, Griffiths PD, Sinzger C, McSharry BP, Wilkinson GW, Davison AJ. 2004. Genetic content of wild-type human cytomegalovirus. *J Gen Virol* 85:1301–1312.
- Hassan-Walker AF, Okwuadi S, Lee L, Griffiths PD, Emery VC. 2004. Sequence variability of the α -chemokine UL146 from clinical strains of human cytomegalovirus. *J Med Virol* 74:573–579.
- He R, Ruan Q, Qi Y, Ma YP, Huang YJ, Sun ZR, Ji YH. 2006. Sequence variability of human cytomegalovirus UL146 and UL147 genes in low-passage clinical isolates. *Intervirology* 49:215–223.
- Iwasenko JM, Scott GM, Ziegler J, Rawlinson WD. 2007. Emergence and persistence of multiple antiviral resistant CMV strains in a highly immunocompromised child. *J Clin Virol* 40:152–155.
- Kall L, Krogh A, Sonnhammer EL. 2004. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 338:1027–1036.
- Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511–518.
- Lurain NS, Fox AM, Lichy HM, Bhorade SM, Ware CF, Huang DD, Kwan SP, Garrity ER, 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.
- Mattick C, Dewin D, Polley S, Sevilla-Reyes E, Pignatelli S, Rawlinson W, Wilkinson G, Dal Monte P, Gompels UA. 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. *J Virol* 318:582–597.
- McGeoch DJ, Rixon FJ, Davison AJ. 2006. Topics in herpesvirus genomics and evolution. *Virus Res* 117:90–104.
- Murphy E, Yu D, Grimwood J, Schmutz J, Dickson M, Jarvis MA, Hahn G, Nelson JA, Myers RM, Shenk TE. 2003. Coding potential of

- laboratory and clinical strains of human cytomegalovirus. *Proc Natl Acad Sci USA* 100:14976–14981.
- Penfold ME, Dairaghi DJ, Duke GM, Saederup N, Mocarski ES, Kemble GW, Schall TJ. 1999. Cytomegalovirus encodes a potent α chemokine. *Proc Natl Acad Sci USA* 96:9839–9844.
- Pignatelli S, Dal Monte P, Landini MP. 2001. gpUL73 (gN) genomic variants of human cytomegalovirus isolates are clustered into four distinct genotypes. *J Gen Virol* 82:2777–2784.
- Pignatelli S, Dal Monte P, Rossini G, Chou S, Gojobori T, Hanada K, Guo JJ, Rawlinson W, Britt W, Mach M, Landini MP. 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* 84:647–655.
- Pignatelli S, Dal Monte P, Rossini G, Landini MP. 2004. Genetic polymorphisms among human cytomegalovirus (HCMV) wild-type strains. *Rev Med Virol* 14:383–410.
- Poncet C, Frances V, Gristina R, Scheiner C, Pellissier JF, Figarella-Branger D. 1996. CD24, a glycosylphosphatidylinositol-anchored molecule, is transiently expressed during the development of human central nervous system and is a marker of human neural cell lineage tumors. *Acta Neuropathol* 91:400–408.
- Prichard MN, Penfold MET, Duke GM, Spaete RR, Kemble GW. 2001. A review of genetic differences between limited and extensively passaged human cytomegalovirus strains. *Rev Med Virol* 11:191–200.
- Pride DT. 2004. Svaap 1.0.1: A tool for analyzing substitutions and similarity in multiple alignments. <http://www.bacteriamuseum.org/SWAAP/SvaapPage.htm>.
- Puchhammer-Stöckl E, Görzer I. 2006. Cytomegalovirus and Epstein-Barr virus subtypes—The search for clinical significance. *J Clin Virol* 36:239–248.
- Puchhammer-Stöckl E, Görzer I, Zoufaly A, Jaksch P, Bauer CC, Klepetko W, Popow-Kraupp T. 2006. Emergence of multiple cytomegalovirus strains in blood and lung of lung transplant recipients. *Transplantation* 81:187–194.
- Qi Y, Mao ZQ, Ruan Q, He R, Ma YP, Sun ZR, Ji YH, Huang Y. 2006. Human cytomegalovirus (HCMV) UL139 open reading frame: Sequence variants are clustered into three major genotypes. *J Med Virol* 78:517–522.
- Rasmussen L, Geissler A, Cowan C, Chase A, Winters M. 2002. The genes encoding the gCIII complex of human cytomegalovirus exist in highly diverse combinations in clinical isolates. *J Virol* 76:10841–10848.
- Rasmussen L, Geissler A, Winters M. 2003. Inter- and intragenic variations complicate molecular epidemiology of human cytomegalovirus. *J Infect Dis* 187:809–819.
- Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.
- Schabath H, Runz S, Joumaa S, Altevogt P. 2006. CD24 affects CXCR4 function in pre-B lymphocytes and breast carcinoma cells. *J Cell Sci* 119:314–325.
- Smith SC, Oxford G, Wu Z, Nitz MD, Conaway M, Frierson HF, Hampton G, 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* 66:1917–1922.
- Staden R, Beal KF, Bonfield JK. 2000. The Staden package, 1998. *Methods Mol Biol* 132:115–130.
- Stanton R, Westmoreland D, Fox JD, Davison AJ, Wilkinson GW. 2005. Stability of human cytomegalovirus genotypes in persistently infected renal transplant recipients. *J Med Virol* 75:42–46.
- Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596–1599.
- Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequencing alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
- Yang Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl BioSci* 13:555–556.