Major article

Multiple-Strain Infections of Human Cytomegalovirus with High Genomic Diversity are Common In Breast Milk from HIV-Positive Women in Zambia

Nicolás M. Suárez[1a], Kunda G. Musonda[2,3a], Eric Escriva[2,4], Margaret Njenga[2b], Anthony Agbueze[2,4], Salvatore Camiolo[1], Andrew J. Davison[1], and Ursula A. Gompels[2b]

[1]MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom; [2]Pathogen Molecular Biology Department, London School of Hygiene and Tropical Medicine, London, United Kingdom; [3]Virology Laboratory, University Teaching Hospital, Lusaka, Zambia; [4]Birkbeck College, University of London, London, United Kingdom

Running title: HCMV diversity in milk

Abstract: 200 words

Main text: 3500 words

A preprint version of this paper was published in December 2018 (BioRxiv 493742. doi: https://doi.org/10.1101/493742).

[a]N. M. S. and K. G. M. contributed equally to this study.

[b]Present affiliation: Public Health England, Porton Down, Wiltshire, United Kingdom.

Correspondence: U. A. Gompels; Pathogen Molecular Biology Department, London School of Hygiene and Tropical Medicine, University of London, Keppel Street, WC1E 7HT, United Kingdom (ursula.gompels@lshtm.ac.uk).

ABSTRACT

Background. In developed countries, human cytomegalovirus (HCMV) is a major pathogen in congenitally infected and immunocompromised individuals, in whom multiple-strain infection is linked to disease severity. The situation is less documente in developing countries. In Zambia, breast milk is a key route for transmitting HCMV and carries higher viral loads in HIV-positive women. We investigated HCMV strain diversity.

Methods. High-throughput sequence datasets were generated from 28 HCMV-positive breast milk samples donated by 22 mothers (15 HIV-positive and seven HIV-negative) at 4 or 16 weeks (or both) postpartum and analysed by genotyping 12 hypervariable HCMV genes.

Results. Among the 20 samples from 14 donors (13 HIV-positive and one HIV-negative) that yielded data meeting quality thresholds, 89 of the possible 109 genotypes were detected, and multiple-strain infections involving up to five strains per person were apparent in nine HIV-positive women. Strain diversity was extensive among individuals but conserved compartmentally and longitudinally within them. Genotypic linkage was maintained within the hypervariable UL73/UL74 and RL12/RL13/UL1 loci for virus-entry and immunomodulation, but not between genes more distant from each other.

Conclusions. Breast milk from HIV-positive women contains multiple HCMV strains of high genotypic complexity and thus constitutes a major source for transmitting viral diversity.

Keywords. Human cytomegalovirus; high-throughput sequencing; breast milk; target enrichment; viral genomics.

52 **BACKGROUND**

53

54 Human cytomegalovirus (HCMV) is a major coinfection in HIV-positive people, in whom, as in
55 other immunocompromised individuals such as transplant recipients, it contributes to morbidity and
56 mortality. HCMV is also the most frequent congenital infection, causing adverse neurodevelopment,
57 including hearing loss, microcephaly and neonatal morbidity. Postnatal infection generally occurs via
58 milk in breastfeeding populations and is usually asymptomatic. However, it has been linked to
59 morbidity, especially in preterm or underweight infants, and, in recent population studies, to
60 adverse developmental effects, especially in association with HIV exposure in developing countries
61 [1-4]. The most severe HCMV infections in transplant recipients, whether due to primary infection,
62 reinfection or reactivation from latency, can result in severe or end-organ diseases such as retinitis,
63 pneumonitis, hepatitis and enterocolitis [5]. Few studies of HCMV diversity, transmission and
64 epidemiology have been conducted in relation to developing countries, including those having a high
65 burden of endemic HIV.

66 HCMV has a double-stranded DNA genome of 236 kbp containing at least 170 protein-coding
67 genes [6]. Diversity among strains is low overall, except in several hypervariable genes that exist as
68 distinct, stable genotypes. These genes encode proteins that are particularly vulnerable to immune
69 selection, including virus entry glycoproteins, other membrane glycoproteins and secreted proteins.
70 The recombinant nature of HCMV strains was first identified in serological surveys and then in
71 genomic studies, and is a key consideration for vaccine development [7-16]. However,
72 understanding the pathogenic effects of HCMV diversity is at an early stage [17-19], and is limited by
73 the fact that most analyses have focused on a few hypervariable genes characterised by polymerase
74 chain reaction (PCR)-based genotyping [7, 12, 20]. This approach is relatively insensitive to the
75 presence of minor strains in multiple-strain infections, which may have more serious outcomes.

76 High-throughput sequencing studies at the whole-genome level have started to facilitate a
77 broader view of HCMV diversity, but most have involved isolating the virus in cell culture, which is

78 prone to strain loss or mutation, or have depended on direct sequencing of PCR amplicons
79 generated from clinical samples [11, 14, 16, 21]. Recent studies have avoided these limitations by
80 using target enrichment to enable direct sequencing of strains present in clinical samples, most of
81 which originated from patients in developed countries with congenital or transplantation-associated
82 infections [11, 22-24]. We have used this approach to examine HCMV strain diversity in a developing
83 country by analysing breast milk from women in Zambia, who constitute an HIV endemic population
84 in sub-Saharan Africa, where we have previously demonstrated the negative developmental effects
85 of early infection of infants with HCMV, particularly alongside HIV exposure [1, 3].

86 **METHODS**

87

88 **Patients and samples**

89 Anonymised breast milk samples were collected with informed consent as a substudy of the
90 Breast Feeding and Postpartum Health study conducted at the University Teaching Hospital, Lusaka,
91 Zambia, as approved by the ethical committees of the University Teaching Hospital and the London
92 School of Hygiene and Tropical Medicine. This substudy included 28 HCMV-positive breast milk
93 samples donated from one or both breasts by 15 HIV-positive and seven HIV-negative mothers at 4
94 or 16 weeks (or both) postpartum (**Table 1** and **Supplementary Table 1** rows 3-6) [3].

95

96 **DNA extraction and viral load quantification**

97 DNA was extracted from 200 µl breast milk using a QIAamp® DNA mini kit (QIAGEN, Manchester,
98 UK), and viral DNA load was measured using an in-house HCMV gB TaqMan assay run on an Applied
99 Biosystems® 7500 fast real-time PCR system (Applied Biosystems, Foster City, CA, USA), as described
100 previously (**Table 1** and **Supplementary Table 1** row 7) [3].

101

102 **High-throughput DNA sequencing**

103 The SureSelect^{XT} v. 1.7 target enrichment system (Agilent, Stockport, UK) was used to prepare
104 sequencing libraries, as described previously (**Supplementary Table 1** rows 8-10) [22]. The libraries
105 were sequenced using a MiSeq (Illumina, San Diego, CA, USA) with v. 3 chemistry to generate
106 original datasets consisting of paired-end reads of 300 nucleotides (nt; **Table 1** and **Supplementary**
107 **Table 1** rows 11-12).

108

109 **Phylogenetic analysis**

110 UL73 and UL74 genotypes [7, 12] were investigated in 243 different HCMV strains [24]. MEGA
111 6.06 [25] was used to generate Muscle-derived amino acid sequence alignments and phylogenetic

112 trees based on the Jones-Taylor-Thornton model and discrete gamma distribution with five
113 categories.
114
115 **Strain characterisation using sequence motifs**
116 Original datasets were quality-checked and trimmed using Trim Galore
117 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/; length=21, quality=10 and
118 stringency=3) (**Supplementary Table 1** row 13). Bowtie2 [26] was used to remove reads mapping to
119 the Genome Reference Consortium Human Reference 38 sequence, and also quality-checked and
120 trimmed to create purged datasets (**Supplementary Table 1** row 14). Dataset quality parameters
121 were (**Supplementary Table 1** rows 19-23) [24] on the basis of thresholds described in Results.
122 The number of genotypes was analysed by counting reads containing conserved, genotype-
123 specific sequence motifs or their reverse complements. One short motif (14 nt) for each UL73
124 genotype and three short motifs (12-13 nt) for each UL74 genotype were identified by examining
125 nucleotide sequence alignments and polymorphism plots derived from the 163 HCMV genome
126 sequences in GenBank Release 211 (15 December 2015). Motif conservation was confirmed in the
127 243 genome set plus 383 UL73 and 72 UL74 single-gene sequences available in GenBank. The
128 sequences of the short motifs are listed in **Table 2** with their frequency of occurrence.
129 In addition to counting short motifs (**Supplementary Table 1** rows 25-56), long motifs (20-24 nt)
130 at one per genotype were counted in UL73 and UL74 and a further ten hypervariable genes (RL5A,
131 RL6, RL12, RL13, UL1, UL9, UL11, UL120, UL146 and UL139) (**Supplementary Table 1** rows 58-166).
132 Long motifs identifying common mutations in three genes (RL5A, UL111A and US9) were also
133 counted (**Supplementary Table 1** rows 167-174). The derivation of the long motifs as described [24],
134 were used on the basis of thresholds described in Results (**Supplementary Table 1** row 17).
135
136 **Variant analysis**

137 Replacement of one strain by another as the major population (genotype switch) in
138 compartmental or longitudinal samples from the same individual were investigated by variant
139 analysis [21, 27]. The original datasets were quality checked using FASTQC
140 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), trimmed to over 100 nt using
141 Trimmomatic [28], optimised using VelvetOptimiser parameters
142 (http://www.vicbioinformatics.com/software.velvetoptimiser.shtml), with assembly *de novo* using
143 Velvet [29] and produced contigs checked by reference genome mapping with ABACAS [30]. The
144 resulting contigs were verified by reference mapping using BWA [31] and SAMtools/BCFtools [32].
145 GATK [33] was used for indexing, mapping and variant calling, defining variant nucleotides as
146 follows: prevalence <50%, overall read depth ≥50, average nucleotide quality ≥30, variant frequency
147 ≥1% for read depths >1000 and >10% for read depths 50-1000, and minimum SNP depth ≥10.
148 Artemis [34] was used for visualisation.
149
150 **Data deposition**
151 The human purged datasets were deposited in the European Nucleotide Archive (ENA) under
152 project number PRJEB31143 (**Supplementary Table 1** row 15). Complete genome sequences were
153 assembled as described [22] and deposited in GenBank (**Supplementary Table 1** row 16).

154 **RESULTS**

155

156 **Assessment of the sequence datasets**

157 In a recent study, we highlighted the importance of monitoring the quality of datasets produced
158 directly from clinical material by target enrichment and high-throughput sequencing [24]. We
159 implemented this here by assembling the datasets against the reference strain Merlin genome
160 (GenBank accession AY446894), noting the number of matching HCMV reads (**Supplementary Table**
161 **1** line 19), and deriving two parameters: (1) the percentage of matching HCMV reads in the dataset
162 and (2) the percentage of the reference genome represented (**Supplementary Table 1** lines 20-21).
163 Also, since sequencing methodology is highly PCR-based, the number of HCMV fragments from
164 which the data were produced was monitored by additional parameters: (3) the coverage depth of
165 the reference genome by all HCMV reads, and (4) the coverage depth of the reference genome by
166 reads generated from unique HCMV DNA fragments (**Supplementary Table 1** lines 22-23).

167 Quality threshold values were set at: (1) ≥50%, (2) ≥95%, (3) ≥1000 of the total fragment
168 reads/nt, and (4) ≥20 unique fragment reads/nt. Eighteen datasets generated from 13 women met
169 all four criteria, and two datasets (141R16 and 243L16) met criteria (1) to (3) but exhibited lower
170 values (between 10-20) for criterion (4). These 20 datasets (one from an HIV-negative women and
171 19 from 13 HIV-positive women) are indicated in **Table 1** and **Supplementary Table 1** row 11, and
172 further analysed.

173

174 **Genotypic structure of the UL73/UL74 locus**

175 Our previous study involving Sanger sequencing of single HCMV genes in breast milk samples
176 obtained at multiple time points postpartum pointed to the presence of multiple strains [3]. We
177 extended this by using sequence differences between the genotypes of hypervariable genes across
178 the genome to characterise the strains represented in the datasets. We focused first on UL73 and
179 UL74, as our earlier work had shown that these adjacent genes are markedly hypervariable, are

180 almost always genotypically linked, and group into eight genotypes also identified in milk samples
181 [3,7,12]. The nucleotide sequences were extracted from the set of 243 genome sequences and
182 analysed phylogenetically (**Figure 1**). This confirmed the existence of eight genotypes for each gene
183 (**Table 2**), the strong linkage between them (only seven recombinants were noted), and the high
184 levels of intergenotypic diversity and low levels of intragenotypic diversity observed initially in small
185 datasets [7, 12]. In the UL73 phylogeny, a single G4B strain (HAN; GenBank accession KJ426589) fell
186 outside the genotypes due to three nucleotide differences that are characteristic of G4A strains and
187 probably represent homoplasies. In the UL74 phylogeny, a single strain (BE/23/2010; GenBank
188 accession KP745697) fell outside the genotypes and may have resulted from intragenic
189 recombination between G1C and G1A. The distances between genotypes and branching in the two
190 phylogenies also supported our previous inference that an ancestral recombination event had given
191 rise to the linkage between UL73 G4C and UL74 G1C [12].
192
193 **Genotyping using sequence motifs**
194 Having established a comprehensive view of UL73 and UL74 hypervariation, we developed short
195 motifs capable of identifying individual genotypes. These consisted of a single motif near the 5' end
196 of each UL73 genotype and three separate motifs near the 5' and 3' ends and in the central region of
197 each UL74 genotype. These motifs genotyped successfully the majority of the sequences used in the
198 phylogenetic analyses (**Table 2**). We then extended the analysis to a further ten hypervariable genes,
199 using a single, long, nonredundant motif for each genotype to improve discrimination.
200 The original datasets were trimmed (to create trimmed datasets) or purged of human reads and
201 trimmed (to create purged datasets). The relative frequencies of individual genotypes were then
202 estimated by counting motifs in all datasets with threshold requirements applied (**Supplementary**
203 **Table 1** lines 25-56 and 58-166, respectively). Exclusion of human reads had little effect, except
204 when short motifs were used with datasets containing a higher proportion of human reads. The
205 UL74 5' motif offered the least accurate genotypic discernment in such samples, perhaps as a result

206 of its minimal length (12 nt). The number of strains in each sample was scored from the purged
207 datasets using the long motifs with threshold requirements (**Table 3** and **Supplementary Table 1**
208 row 17). A genotype was considered to be present when represented by >25 reads and >5% of the
209 total number of reads detected for all genotypes of that gene, and the number of strains was scored
210 as being the greatest number of genotypes detected using long motifs for at least two genes. Thus,
211 strains present at <5% were unlikely to have been scored. There was a high degree of congruence
212 between the results obtained using short and long motifs with datasets meeting the quality
213 thresholds (**Supplementary Table 2**).
214
215 **Strain complexity in HIV-positive women**
216 The majority of HIV-positive women (11/13) were infected by multiple HCMV strains (**Table 3**
217 and **Supplementary Table 1**). The mode was at least two strains, and one woman was infected by
218 five strains. The only HIV-negative woman was infected by a single strain, but it was not possible to
219 discern from this whether multiple strains are more common in HIV-positive women. (The data from
220 the other six HIV-negative women that were excluded from the analysis also indicate single
221 infections in this group, but failure to meet the quality thresholds, partially from lower viral loads,
222 could lead to underestimations). Even among this small cohort, 89 of the 109 possible genotypes of
223 the 12 genes were detected. It was possible to assign fully linked genotypes (haplotypes) to eight
224 strains represented in 11 datasets from seven donors, on the basis of complete genome sequences
225 (four datasets) or the presence of a single strain or major and minor strains (when the former was
226 highly predominant) in multiple-strain infections (**Table 3**). Consideration of other datasets indicated
227 partially linked genotypes for up to 30 strains (**Supplementary Table 3**).
228 Genotypic linkage was detected only in two loci within which recombination has been shown to
229 occur rarely, namely those containing the two respective sets of adjacent, hypervariable genes
230 UL73/UL74 [12, 17, 35] and RL12/RL13/UL1 [11, 16]. The overall frequencies of UL73/UL74
231 genotypes in the milk samples were not significantly different from those in the 243 genome set plus

232 single-gene sequences (**Tables 1** and **3**). There was some evidence for increased proportions of
233 UL73/74 G4B-G4 and RL12/RL13/UL1 G2-G2-G2 in milk (p=.001 and p=0.02 respectively, compared
234 to the 243 genome set) , but case-controlled cohorts are required to confirm.

235 The use of three short motifs in UL74 facilitated an examination of intragenic recombination,
236 and confirmed that strain BE/23/2010 is a recombinant with a G1C motif near the 5' end and G1A
237 motifs in the central region and near the 3' end. In addition, compartmental stability was revealed
238 by the use of both short and long motifs, in the form of genotypic conservation in samples from both
239 breasts of four HIV-positive women (**Figure 2**). Small differences may be accounted for by minor
240 strains present at levels nearing the detection threshold. Longitudinal stability was observed in two
241 donors (243 and 278) with samples taken at weeks 4 and 16 postpartum (**Table 3**); small differences
242 in one (243) were probably from threshold effects. This stability also showed in variant analysis,
243 which demonstrated the absence of genotype switches in all donors. Such switches may be caused
244 by the differential growth advantages of individual strains or intrahost recombinants, as reported
245 recently in transplant recipients [22].

246 Finally, additional long sequence motifs were used to investigate whether any strains contained
247 gene-disrupting mutations detected previously in the set of 243 genomes [24]. These virus
248 'pseudogene' mutations cause frameshifting, resulting in loss or modification of the encoded
249 proteins. They are more common in certain genes, most frequently and in order of decreasing
250 abundance UL9, RL5A, UL1, RL6, US9 and UL111A [14, 16, 24]. The use of motifs representing three
251 mutations in RL5A (present in 37 members of the 243 genome set), two in US9 (35 members) and
252 one in UL111A (five members) demonstrated the presence of the RL5A and US9 mutations, but not
253 that in UL111A, encoding viral interleukin-10 (**Supplementary Table 1** rows 167-174).

254 **DISCUSSION**

255 Analysis of HCMV genomes directly from clinical samples is necessary for characterising
256 infectious natural populations while avoiding the mutational artefacts arising from laboratory
257 adaptation to cell-culture. Target enrichment has proven successful in this regard [11, 22, 24], but
258 accurate genome analysis can be confounded by multiple strains, particularly in immunosuppressed
259 groups in whom additional complexity may accumulate by reinfection or reactivation [17, 21, 24].
260 We have shown previously that HIV-positive women in sub-Saharan Africa have higher HCMV loads
261 in breast milk than HIV-negative women, and that this is associated with adverse infant
262 development [1, 3]. However, genomic studies of HCMV in milk samples or, indeed, samples from
263 Africa, are scarce. We examined milk because of its importance in HCMV transmission, with the aim
264 of understanding strain diversity and the burden of infection in HIV-positive (immunosuppressed)
265 mothers, which may affect their infants. The sequence datasets were generated from 28 samples
266 donated by 22 women, and 20 datasets from 14 women meeting quality thresholds were analysed.

267 The analysis focused on counting reads containing motifs specific to the genotypes of
268 hypervariable genes. Short motifs were developed initially for sensitive characterisation of UL73 and
269 UL74, which encode glycoproteins N and O (gN and gO), respectively, and then long motifs were
270 used for further resolution of these two genes and ten others. Since UL73 and UL74 are linked and
271 behave as a single genotype, as further shown here, haplotypes could not be determined using
272 solely the short motifs (**Supplementary Table 2**) [7, 12]. However, mapping three short motifs to
273 each UL74 genotype was uniquely useful for detecting intragenic recombination. The use of long
274 motifs in a larger number of genes allowed increased resolution and haplotype information. These
275 were less compromised by residual human reads in the datasets, but more susceptible to
276 mismatches in target genomes (**Table 3**).

277 Genotypic and haplotypic complexity in this small cohort was remarkable. Most (82%) of the
278 genotypes possible in the 12 hypervariable genes were detected, and 85% of the HIV-positive donors
279 were infected by multiple strains. The level of multiple strain infection exceeds that in previous

280 cohort analyses, including congenitally-infected and transplantation patients from developed
281 countries, 243 set [22,24]. Each of the eight fully characterised haplotypes identified was unique in
282 this cohort and also in the set of 243 strains, in which most strains (223) are also unique [24]. In the
283 milk cohort, up to 30 strains could be determined, using partial haplotypes. These observations
284 testify to the huge number of HCMV haplotypes that may exist, possibly exceeding that related to
285 immune diversity, as was recognised long before the high-throughput sequencing era [7, 8, 9, 12,
286 13,15]. No evidence emerged for the existence of novel African genotypes, consistent with the view
287 that HCMV genotypes are distributed throughout the world, although their relative prevalence may
288 vary [7, 8, 12].

289 Strain composition in individual women was essentially stable, compartmentally (in milk samples
290 from both breasts) and longitudinally (at 4 and 16 weeks postpartum). This indicates that the strains
291 detected were present in the donor prior to viral reactivation in breast tissue during lactation.
292 Leukocyte infiltrates have been characterised during this period [36], and may be the source of
293 reactivated virus. A study conducted in Uganda has shown that mothers can be infected with strains
294 from children [37]. However, even though multiple-strain infections were common in the cohort,
295 there was limited evidence for reinfection or reactivation during the 4-16 week period postpartum.
296 It is likely that opportunities for fresh infection existed at home and in the hospital, because all the
297 mothers were HCMV-positive and had young children at home. The high proportion of multiple-
298 strain infections observed also provides in principle for the diversification by recombination. These
299 observations differ from those made in developed countries. A proportion of transplant-associated
300 infections involve multiple strains, and these exhibit substantial longitudinal dynamism [22, 24] and
301 are also associated with increased viral loads and the pathological outcomes of HCMV disease [18,
302 20]. The contrasting observation that most congenital or postnatal infections involve single strains
303 [24, 38] suggests that only certain strains cross the placenta or are transmitted by breast milk, urine
304 or saliva, perhaps due to the competence of a few virions to establish infection [38, 39]. This also
305 implies that the HIV-positive women were exposed to a high burden of superinfection.

306 Whole-genome analyses and earlier PCR-based studies showed a high degree of linkage within
307 the UL73/UL74 [7,12, 24,35] and RL12/RL13/UL1 loci [11, 24]. This is consistent with the
308 involvement of homologous recombination during HCMV evolution, and may also reflect the
309 functional constraints imposed on proteins that interact with each other or have interdependent
310 functions. The UL73 and UL74 proteins (gN and gO) are part of the viral entry complex and have
311 roles in viral exocytosis, cellular tropism and modulation of antibody neutralisation, and the RL12,
312 RL13 and UL1 proteins are known or suspected to be involved in aspects of immune evasion
313 probably mediated by an immunoglobulin-like binding domain shared by these proteins and other
314 members of the RL11 family [9, 13, 40-43]. In addition, RL13 may influence the effect of UL74 on the
315 growth of HCMV [44]. It is possible that different genotypes of hypervariable genes, and different
316 combinations of genotypes, provide variable growth properties leading to higher viral loads and
317 specific pathologies. For example, UL74 genotypes differentially affects viral growth properties *in*
318 *vitro* [45], and genotypes of UL146, which is the most hypervariable gene in HCMV and encodes a
319 vCXCL1 chemokine, affects neutrophil chemotaxis efficiency [46]. Similarly, host variation would be
320 higher in Africa and may affect susceptibility as a result of HCMV genotype-specific interactions, for
321 example with immunoglobulin variants [47-50].
322 Although data on genotypes and mutants could be extracted from the datasets regardless of
323 strain complexity, complete genome sequence determination was possible for only four datasets
324 because of high frequency of confounding multiple infections. To our knowledge, these are the first
325 complete HCMV genome sequences to be determined from people living in Africa. Moreover, one of
326 these originated from an HIV-negative woman and thus represents the first from an
327 immunocompetent adult lacking HCMV-associated pathology. Future research is likely to focus on
328 understanding the differences in HCMV transmission in immunosuppressed and immunocompetent
329 settings in order to define the interplay between viral strain and host immunotype diversity in
330 controlling disease.

331 **NOTES**

332

333 **Acknowledgements.** The authors thank the participants and clinical staff, managed by Dr L. Kasonka
334 at the University Teaching Hospital, Lusaka, for facilitating the Breastfeeding and Postpartum Health
335 Study, directed by Professor S. Filteau, London School of Hygiene and Tropical Medicine, for enabling
336 the follow up analyses. We are grateful to Dr J. Hughes, MRC-University of Glasgow Centre for Virus
337 Research, for advice on bioinformatic analysis, and Drs T. Clark and J. Phelan, London School of
338 Hygiene and Tropical Medicine, for facilitating UNIX cluster access and Perl support.

339

340 **Supplementary data.** Supplementary materials are available at The Journal of Infectious Diseases
341 online. These consist of data provided by the authors to benefit the reader, are not copyedited, and
342 are the sole responsibility of the authors. Questions or comments should be addressed to the
343 corresponding author.

344

345 **Financial support.** This work was supported by the Commonwealth Scholarship Commission, a
346 Bloomsbury Studentship Award and the Medical Research Council (grant MC_UU_12014/3).

347

348 **Potential conflicts of interest.**The authors report no conflicts of interest in this study.
349 All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

350 **REFERENCES**

351

352 1. Gompels UA, Larke N, Sanz-Ramos M, et al. Human cytomegalovirus infant infection adversely
353 affects growth and development in maternally HIV-exposed and unexposed infants in Zambia.
354 Clin Infect Dis **2012**; 54:434-42.
355 2. Josephson CD, Caliendo AM, Easley KA, et al. Blood transfusion and breast milk transmission of
356 cytomegalovirus in very low-birth-weight infants: a prospective cohort study. JAMA Pediatr
357 **2014**; 168:1054-62.
358 3. Musonda KG, Nyonda M, Filteau S, Kasonka L, Monze M, Gompels UA. Increased
359 cytomegalovirus secretion and risks of infant infection by breastfeeding duration from maternal
360 human immunodeficiency virus positive compared to negative mothers in sub-Saharan Africa. J
361 Pediatric Infect Dis Soc **2016**; 5:138-46.
362 4. Hamprecht K, Goelz R. Postnatal cytomegalovirus infection through human milk in preterm
363 infants: transmission, clinical presentation, and prevention. Clin Perinatol **2017**; 44:121-130.
364 5. Griffiths P, Baraniak I, Reeves M. The pathogenesis of human cytomegalovirus. J Pathol **2015**;
365 235:288-97.
366 6. Gatherer D, Seirafian S, Cunningham C, et al. High-resolution human cytomegalovirus
367 transcriptome. Proc Natl Acad Sci U S A **2011**; 108:19755-60.
368 7. Bates M, Monze M, Bima H, et al. High human cytomegalovirus loads and diverse linked variable
369 genotypes in both HIV-1 infected and exposed, but uninfected, children in Africa. Virology **2008**;
370 382:28-36.
371 8. Bradley AJ, Kovacs IJ, Gatherer D, et al. Genotypic analysis of two hypervariable human
372 cytomegalovirus genes. J Med Virol **2008**; 80:1615-23.
373 9. Davison AJ, Akter P, Cunningham C, et al. Homology between the human cytomegalovirus RL11
374 gene family and human adenovirus E3 genes. J Gen Virol **2003**; 84:657-63.

375 10. Dolan A, Cunningham C, Hector RD, et al. Genetic content of wild-type human cytomegalovirus. J
376 Gen Virol **2004**; 85:1301-12.
377 11. Lassalle F, Depledge D, Reeves MB, et al. Islands of linkage in an ocean of pervasive
378 recombination reveals two-speed evolution of human cytomegalovirus genomes. Virus Evol
379 **2016**; 2:vew017.
380 12. Mattick C, Dewin D, Polley S, et al. Linkage of human cytomegalovirus glycoprotein gO variant
381 groups identified from worldwide clinical isolates with gN genotypes, implications for disease
382 associations and evidence for N-terminal sites of positive selection. Virology **2004**; 318:582-97.
383 13. Paterson DA, Dyer AP, Milne RS, Sevilla-Reyes E and Gompels UA. A role for human
384 cytomegalovirus glycoprotein O (gO) in cell fusion and a new hypervariable locus. Virology **2002**;
385 293:281-94.
386 14. Cunningham C, Gatherer D, Hilfrich B, et al. Sequences of complete human cytomegalovirus
387 genomes from infected cell cultures and clinical specimens. J Gen Virol **2010**; 91:605-15.
388 15. Rasmussen L, Geissler A, Winters M. Inter- and intragenic variations complicate the molecular
389 epidemiology of human cytomegalovirus. J Infect Dis **2003**; 187:809-19.
390 16. Sijmons S, Thys K, Mbong Ngwese M, et al. High-throughput analysis of human cytomegalovirus
391 genome diversity highlights the widespread occurrence of gene-disrupting mutations and
392 pervasive recombination. J Virol **2015**; 89:7673-95.
393 17. Gorzer I, Guelly C, Trajanoski S, Puchhammer-Stockl E. Deep sequencing reveals highly complex
394 dynamics of human cytomegalovirus genotypes in transplant patients over time. J Virol **2010**;
395 84:7195-203
396 18. Puchhammer-Stockl E, Gorzer I, Zoufaly A, et al. Emergence of multiple cytomegalovirus strains
397 in blood and lung of lung transplant recipients. Transplantation **2006**; 81:187-94
398 19. Ross SA, Arora N, Novak Z, Fowler KB, Britt WJ, Boppana SB. Cytomegalovirus reinfections in
399 healthy seroimmune women. J Infect Dis **2010**; 201:386-9

400 20. Gorzer I, Kerschner H, Jaksch P, et al. Virus load dynamics of individual CMV-genotypes in lung
401 transplant recipients with mixed-genotype infections. J Med Virol **2008**; 80:1405-14.
402 21. Renzette N, Gibson L, Bhattacharjee B, et al. Rapid intrahost evolution of human
403 cytomegalovirus is shaped by demography and positive selection. PLoS Genet **2013**;9:e1003735.
404 22. Hage E, Wilkie GS, Linnenweber-Held S, et al. Characterization of human cytomegalovirus
405 genome diversity in immunocompromised hosts by whole-genome sequencing directly from
406 clinical specimens. J Infect Dis **2017**; 215:1673-83.
407 23. Houldcroft CJ, Bryant JM, Depledge DP, et al. Detection of low frequency multi-drug resistance
408 and novel putative maribavir resistance in immunocompromised pediatric patients with
409 cytomegalovirus. Front Microbiol **2016**; 7:1317.
410 24. Suárez NM, Wilkie GS, Hage E, et al. Human cytomegalovirus genomes sequenced directly from
411 clinical material: variation, multiple-strain infection, recombination and mutation. BioRxiv **2018**;
412 http://biorxiv.org/cgi/content/short/505735v1.
413 25. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics
414 Analysis version 6.0. Mol Biol Evol **2013**; 30:2725-9.
415 26. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods **2012**; 9:357-
416 9.
417 27. Tweedy JG, Escriva E, Topf M, Gompels UA. Analyses of tissue culture adaptation of human
418 herpesvirus-6A by whole genome deep sequencing redefines the reference sequence and
419 identifies virus entry complex changes. Viruses **2018**; 10:16.
420 28. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.
421 Bioinformatics **2014**; 30:2114-20.
422 29. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs.
423 Genome Res **2008**; 18:821-9.
424 30. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic
425 contiguation of assembled sequences. Bioinformatics **2009**; 25:1968-9.

426 31. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
427 Bioinformatics **2009**; 25:1754-60.
428 32. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools.
429 Bioinformatics **2009**; 25:2078-9.
430 33. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for
431 analyzing next-generation DNA sequencing data. Genome Res **2010**; 20:1297-303.
432 34. Rutherford K, Parkhill J, Crook J, et al. Artemis: sequence visualization and annotation.
433 Bioinformatics **2000**; 16:944-5.
434 35. Yan H, Koyano S, Inami Y, et al. Genetic linkage among human cytomegalovirus glycoprotein N
435 (gN) and gO genes, with evidence for recombination from congenitally and post-natally infected
436 Japanese infants. J Gen Virol **2008**; 89:2275-9.
437 36. Maschmann J, Goelz R, Witzel S, et al. Characterization of human breast milk leukocytes and
438 their potential role in cytomegalovirus transmission to newborns. Neonatology **2015**; 107:213-9.
439 37. Boucoiran I, Mayer BT, Krantz EM, et al. Nonprimary maternal cytomegalovirus infection after
440 viral shedding in infants. Pediatr Infect Dis J **2018**; 37:627-31.
441 38. Gorzer I, Trajanoski S, Popow-Kraupp T, Puchhammer-Stockl E. Analysis of human
442 cytomegalovirus strain populations in urine samples of newborns by ultra deep sequencing. J
443 Clin Virol **2015**; 73:101-104.
444 39. Mayer BT, Krantz EM, Swan D, et al. Transient oral human cytomegalovirus infections indicate
445 inefficient viral spread from very few initially infected cells. J Virol **2017**; 91:e00380-17.
446 40. Jiang XJ, Adler B, Sampaio KL, et al. UL74 of human cytomegalovirus contributes to virus release
447 by promoting secondary envelopment of virions. J Virol **2008**; 82:2802-12.
448 41. Kropff B, Burkhardt C, Schott J, et al. Glycoprotein N of human cytomegalovirus protects the
449 virus from neutralizing antibodies. PLoS Pathog **2012**; 8:e1002999.
450 42. Scrivano L, Sinzger C, Nitschko H, Koszinowski UH, Adler B. HCMV spread and cell tropism are
451 determined by distinct virus populations. PLoS Pathog **2011**; 7:e1001256.

452 43. Wu Y, Prager A, Boos S, et al. Human cytomegalovirus glycoprotein complex gH/gL/gO uses
453 PDGFR-alpha as a key for entry. PLoS Pathog **2017**; 13:e1006281.
454 44. Laib Sampaio K, Stegmann C, Brizic I, et al. The contribution of pUL74 to growth of human
455 cytomegalovirus is masked in the presence of RL13 and UL128 expression. J Gen Virol **2016**;
456 97:1917-27.
457 45. Kalser J, Adler B, Mach M, Kropff B, Puchhammer-Stockl E, Gorzer I. Differences in growth
458 properties among two human cytomegalovirus glycoprotein O genotypes. Front Microbiol **2017**;
459 8:1609.
460 46. Heo J, Dogra P, Masi TJ, et al. Novel human cytomegalovirus viral chemokines, vCXCL-1s, display
461 functional selectivity for neutrophil signaling and function. J Immunol **2015**; 195:227-36.
462 47. Corrales-Aguilar E, Trilling M, Hunold K, et al. Human cytomegalovirus Fcgamma binding proteins
463 gp34 and gp68 antagonize Fcgamma receptors I, II and III. PLoS Pathog **2014**; 10:e1004131.
464 48. Cortese M, Calo S, D'Aurizio R, Lilja A, Pacchiani N, Merola M. Recombinant human
465 cytomegalovirus (HCMV) RL13 binds human immunoglobulin G Fc. PLoS One **2012**; 7:e50166.
466 49. Di Bona D, Accardi G, Aiello A, et al. Association between gamma marker, human leucocyte
467 antigens and killer immunoglobulin-like receptors and the natural course of human
468 cytomegalovirus infection: a pilot study performed in a Sicilian population. Immunology **2018**;
469 153:523-31.
470 50. Pandey JP, Namboodiri AM, Mohan S, Nietert PJ, Peterson L. Genetic markers of
471 immunoglobulin G and immunity to cytomegalovirus in patients with breast cancer. Cell
472 Immunol **2017**; 312:67-70.

473 **FIGURE LEGENDS**

474

475 **Figure 1**

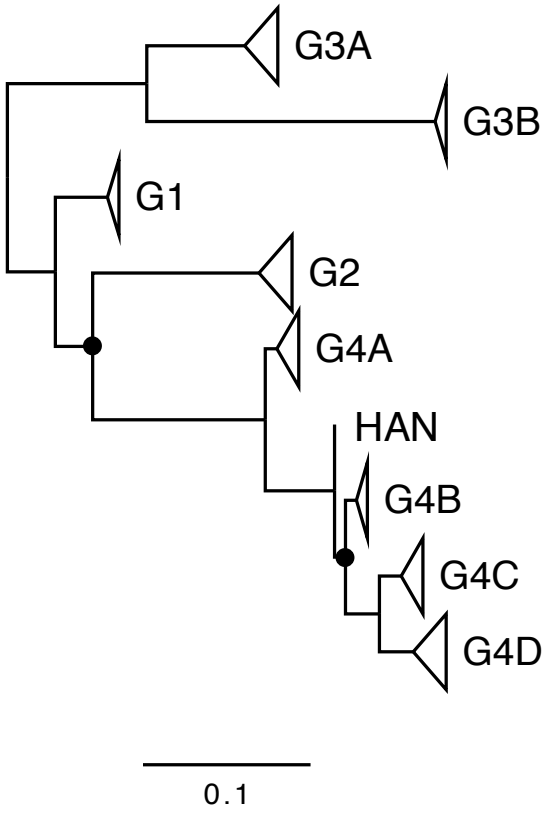
476 Unrooted phylogenetic trees for UL73 and UL74 based on amino acid sequences derived from
477 243 genome sequences, and a summary of genotypic linkages and frequencies. The site coverage
478 cutoff value was 95%, leaving 134 sites in the UL73 tree (log likelihood -1117.87) and 435 sites in the
479 UL74 tree (log likelihood -4153.86). Branch point robustness was inferred from 100 bootstrap
480 replicates, and values of <70% are denoted by filled circles. Genotype branches are collapsed, and
481 the numbers of substitutions per site are shown by the scale. The UL73 sequence of strain HAN and
482 the UL74 sequence of strain BE/23/2010 did not fall into the genotypes. The linkages between UL73
483 and UL74 genotypes are listed, followed by the frequencies of UL73 genotypes in the 243 genome
484 sequences plus single-gene sequences (243+; 626 in total; **Table 2**), and the frequencies of the
485 deduced linkages in the samples (Milk; 26 in total; Table 3). The frequency of each genotype in the
486 Milk set was not significantly different (p=0.05) from that in the 243+ set, as determined by random
487 subsampling analysis (10,000 samplings of 26 genotypes from the set of 626), although there were
488 some differences compared to the 243 genome set as noted in the text. Although no examples of
489 this linkage were present in the datasets at levels in excess of the thresholds, at least one patient
490 (258) was infected at sub-threshold levels by a relevant strain (**Supplementary Table 1**).

491

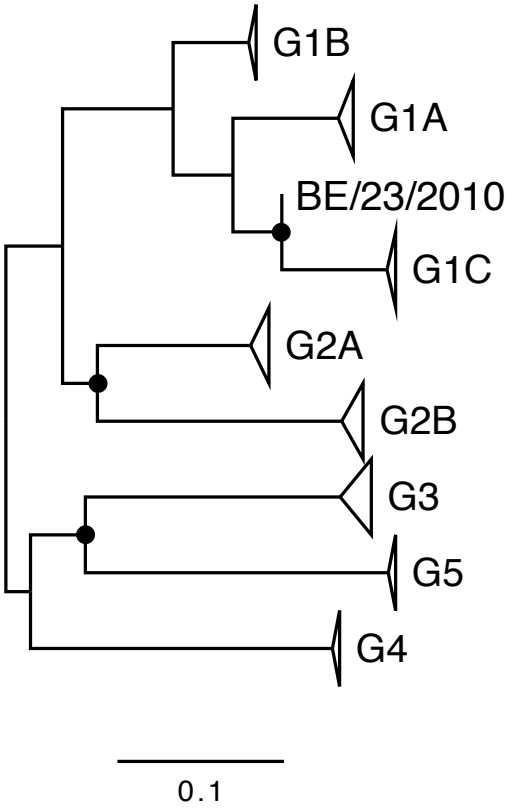
492 **Figure 2**

493 UL73 and UL74 genotypes in milk samples collected from the left (L) and right (R) breasts of four
494 HIV-positive donors at 16 weeks postpartum (**Table 1**). The inner and outer rings show the results
495 obtained using short and long motifs, respectively. Short motif 3' was used for UL74 (**Table 2**). The
496 colour key for genotypes is shown at the foot. Reads that did not meet the inclusion criteria for
497 genotyping are shown as "other".

UL73 (gN)



UL74 (gO)



Linkage and frequency

UL73-UL74	*243+ (%)*	*Milk (%)*
G1-G1A	19	15
G2-G2B	8	12
G3A-G1B	17	23
G3B-G2A	8	4
G4A-G3	18	15
G4B-G4	9	19
G4C-G1C	6	0*
G4D-G5	15	12

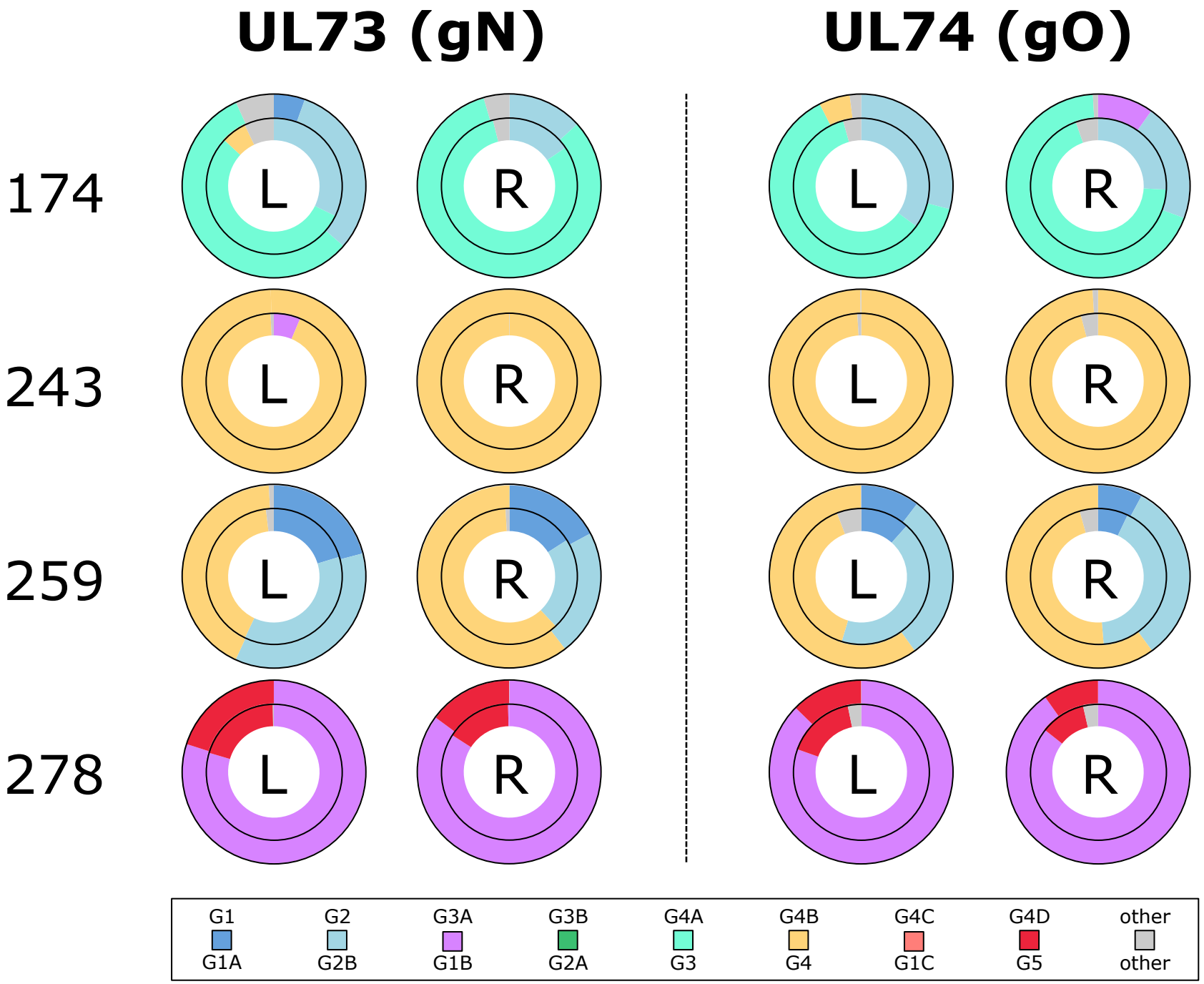

UL73 (gN)
UL74 (gO)
174
243
259
278
L
R
G1
G1A
G2
G2B
G3A
G1B
G3B
G2A
G4A
G3
G4B
G4
G4C
G1C
G4D
G5
other
other

Table 1. Characteristics of donors, samples and datasets.

Donor	HIV status	Breast sample	Weeks postpartum	HCMV load (ge/ml[a])	Dataset
158	Negative	Left	16	818,244	158L16
166	Negative	Left	16	282,252	166L16
193	Negative	Right	16	215,217	193R16[b]
232	Negative	Right	4	470,150	232R4
239	Negative	Right	4	4,752,875	239R4
263	Negative	Left	4	5,285,775	263L4
280	Negative	Right	4	7,505,536	280R4
141	Positive	Right	16	319,888	141R16[c]
154	Positive	Left	16	2,195,856	154L16
173	Positive	Right	16	349,532	173R16[b]
174	Positive	Left	16	371,027	174L16[b]
174	Positive	Right	16	642,516	174R16[b]
181	Positive	Left	4	1,972,365	181L4
243	Positive	Left	16	643,895	243L16[c]
243	Positive	Right	4	65,511,020	243R4[b]
243	Positive	Right	16	795,092	243R16[b]
248	Positive	Right	4	441,679	248R4[b]
258	Positive	Right	4	610,613	258R4[b]
259	Positive	Left	16	2,366,193	259L16[b]
259	Positive	Right	16	5,053,047	259R16[b]
264	Positive	Left	16	388,519	264L16[b]
277	Positive	Right	16	250,377	277R16[b]
278	Positive	Left	16	3,751,776	278L16[b]
278	Positive	Right	4	294,246,272	278R4[b]
278	Positive	Right	16	4,370,800	278R16[b]

281	Positive	Right	4	20,291,530	281R4 [b]
283	Positive	Right	16	274,391	283R16[b]
288	Positive	Right	4	31,574,022	288R4[b]

[a]ge, genome equivalents.

[b]Met all quality thresholds.

[c]Met all quality thresholds except that unique fragment coverage depth was 10-20 rather than ≥20 reads/nt.

Table 2. Short motif sequences in UL73 and UL74.

Gene	Position[a]	Genotype	Motif sequence[b]	Sequences (no.)[c]	Occurrences (no.)[c]	Frequency (%)[c]
UL73	5’	G1	GCGTATCAACTACC	121	121	100
		G2	GTGTGTCGACGAGT	53	53	100
		G3A	GCGTGTCAACAAGC	104	104	100
		G3B	GTGTATCAACGGTA	47	47	100
		G4A	GCACCTTAACAACC	114	113	99
		G4B	ACACCTCAACGACC	55	55	100
		G4C	GCACCTCAACAACC	39	38	97
		G4D	ACGCCTCAACAACC	93	92	99
UL74	5’	G1A	AAACGACWATTT	47	43	91
		G1B	AAAAGGATATCT	60	60	100
		G1C	AAAGGGAACCTT	19	19	100
		G2A	AACCTATTCCTT	27	27	100
		G2B	AGAGCGACATAT	38	38	100
		G3	CGAGCCAGGATT	66	64	97
		G4	AAACAGGTGATT	19	19	100
		G5	TGTCTACATCAT	38	38	100
UL74	C	G1A	CCTTGTGGTACTG	47	47	100
		G1B	TCTTGCGGTACGG	60	60	100
		G1C	TCTTGTGGTACAG	19	19	100
		G2A	TCGTGTGGCGCAG	27	27	100
		G2B	CCTTGCGGTACAG	38	38	100
		G3	TCTTGTGGCACTG	66	66	100
		G4	TCCTGTGGYACGA	19	19	100
		G5	CCTTGYGGCACAG	38	38	100
UL74	3’	G1A	TATTACTACCGCC	47	47	100

G1B	TGTTACTACCACC	60	60	100
G1C	GGTTACCACCAGC	19	19	100
G2A	TGTTACCACCACC	27	27	100
G2B	TGTTACAACCACC	38	38	100
G3	TGCTACCACCACT	66	66	100
G4	TCCTATTGTCCCA	19	19	100
G5	TGCTACCGCTGCT	38	38	100

[a]5', towards the 5' end of the protein-coding region; C, in the central part of the protein-coding region; 3', towards the 3' end of the protein-coding region; in reference strain Merlin, the UL73 5' motif is located at 104-117 nt in UL73 (408 nt; G4D), and the UL74 5', C and 3' motifs are located at 206-217, 443-454 and 906-918 nt, respectively, in UL74 (1419 nt; G5).

[b]UL73 and UL74 are transcribed rightward and leftward, respectively, in the HCMV genome; the sequences are presented 5'-3' in relation to the direction of transcription; IUPAC nucleotide codes are used.

[c]The total number of sequences in the 243 genome set plus single-gene sequences, followed by the number and percentage of these sequences possessing the motif; one UL74 intergenic recombinant (BE/23/2010) was excluded. This provides a measure of motif sensitivity.

Table 3. Genotypes and haplotypes assigned to datasets.

Donor	Dataset	Strains[c]	Genotypes[d]											
			RL5A	RL6	RL12	RL13	UL1	UL9	UL11	UL73	UL74	UL120	UL146	UL139
193	193R16[ab]	1	1	3	8	8	8	4	1	3A	1B	2B	12	3
141	141R16	2	1,6	2,3	2,4B	2	2,4	2,3	5,6	2,3A	1B,2B	4B	2,12	3,8
173	173R16	5	1,2,6	2,3,4	1B,3,4B,6,8	1,6,8	1,6,8	2,4,6,7,9	1,3,5	1,3A,4A,4B	1A,1B,3,4	1A,4B	3,7,9,12	1A,1B,3,7
174	174L16	3	2	2,4	1B,6	1,6,10	1,4,6	4,6	1,4	1,2,4A	2B,3,4	2B,3A	7,9	5,7
174	174R16	3	1,2	1,2,3,4	1B,6	1,6,10	1,6	4,6,7	1,4	2,4A	1B,2B,3	2B,3A	7,9	5,7
243	243L16[a]	2	1	6	2,(10)	(1),2	2	3	6	4B	4	2A	8,9	7
243	243R4[ab]	1	1	6	2	2	2	3	(1),6	4B	4	2A	8	7
243	243R16[a]	1	1	6	2	2	2	3	6	4B	4	2A	8	7
248	248R4[ab]	1	1	1	4A	4A	4	1	1	3B	2A	4B	[5]	4,(7)
258	258R4[a]	2	6	3	3	3	3	3	6	3A,(4A)	1B,(3)	(2A),2B	(2),9	5
259	259L16	3	1,2	2,3,4	1A,6,8	1,6,8	1,6,8	1,3,6	1,4,6	1,2,4B	1A,2B,4	4B	1,9,12	2,8
259	259R16	3	1,2	2,3,4	1A,6,8	1,6,8	1,6,8	1,3,6	1,4,6	1,2,4B	1A,2B,4	4B	1,9,12	2,8
264	264L16	2	1	1,2	8,10	8,10	10	8	4,7	3A,4B	4	1A,3A	10	3
277	277R16	2	1	3	4A,6	4A,6	4,6	6,9	1,4	1,4A	1A,3	3A,4A	1,9	3,4
278	278L16[a]	2	(2),6	3,(4)	(1A),9	(1),[9]	(1),9	(1),9	(1),6	3A,(4D)	1B,(5)	(3A),4B	8,(9)	2,(2)
278	278R4[a]	2	(2),6	3,(4)	(1A),9	(1),[9]	(1),9	(1),9	(1),6	3A,(4D)	1B,(5)	(3A),4B	8,(9)	2,(2)
278	278R16[a]	2	(2),6	3,(4)	(1A),9	(1),[9]	(1),9	(1),9	(1),6	3A,(4D)	1B,(5)	(3A),4B	8,(9)	2,(2)
281	281R4[a]	2	(1),6	3	1A,(6)	1	1,(10)	4,(7)	1	4D	(1B),5	4B	11	5,(7)
283	283R16[ab]	1	[2]	4	4B	[4B]	4	2	5	3A,(3B)	1B	3A	1	5

288	288R4	3	1,2	3,4	6,7	6,8	5,6,8	4,6,9	1	1,4B,4D	1A,4,5	2B,4B	1,3	4,5

[a]Datasets from which haplotypes were assigned.

[b]Datasets from which complete genome sequences were derived.

[c]Determined using long motifs for 12 genes (**Supplementary Table 1**).

[d]Genotype (G) prefix omitted; round brackets indicate an assigned minority genotype; multiple genotypes with none in round parentheses indicate that majority and minority genotypes could not be distinguished; square brackets indicate a single mismatch in the motif.