Major article

Human cytomegalovirus genomes sequenced directly from clinical material: variation, multiple-strain infection, recombination and mutation

Nicolás M. Suárez[1#], Gavin S. Wilkie[1#], Elias Hage[2,3], Salvatore Camiolo[1], Marylouisa Holton[1], Joseph Hughes[1], Maha Maabar[1], Vattipally B. Sreenu[1], Akshay Dhingra[2], Ursula A. Gompels[4], Gavin W. G. Wilkinson[5], Fausto Baldanti[6,7], Milena Furione[6], Daniele Lilleri[8], Alessia Arossa[9], Tina Ganzenmueller[2,3,10], Giuseppe Gerna[8], Petr Hubáček[11], Thomas F. Schulz[2,3], Dana Wolf[12], Maurizio Zavattoni[6] and Andrew J. Davison[1]

[1]MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom; [2]Institute of Virology, Hannover Medical School, Hannover, Germany; [3]German Center for Infection Research, Hannover-Braunschweig Site, Germany; [4]Pathogen Molecular Biology Department, London School of Hygiene and Tropical Medicine, London, United Kingdom; [5]Division of Infection and Immunity, School of Medicine, Cardiff University, Cardiff, United Kingdom; [6]Molecular Virology Unit, Microbiology and Virology Department, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy; [7]Department of Clinical, Surgical, Diagnostic and Pediatric Sciences, University of Pavia, Pavia, Italy; [8]Laboratory of Genetics-Transplantology and Cardiovascular Diseases, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy; [9]Departments of Obstetrics and Gynecology, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy; [10]Institute for Medical Virology and Epidemiology of Viral Diseases, University Hospital Tuebingen, Tuebingen, Germany; [11]Department of Medical Microbiology, Motol University Hospital, Prague, Czech Republic; [12]Clinical Virology Unit, Department of Clinical

Microbiology and Infectious Diseases, Hadassah University Hospital, Jerusalem, Israel

Current addresses: Marylouisa Holton, SGS Vitrology Ltd., 5 South Avenue, Clydebank Business Park, Glasgow G81 2LG, United Kingdom; Maha Maabar, IT Services – Business Systems Team, Level 6, Gilbert Scott Building, University of Glasgow, University Avenue, Glasgow G12 8QQ, United Kingdom; Gavin Wilkie, Illumina, 3/1 International Court, Scoreseby, VIC 3179, Australia

#These authors contributed equally

Running title: Clinical HCMV genomes

Abstract word count: 149

Text word count: 3503

Figures and Tables: 4

Supplementary Tables: 6

References: 50

FOOTNOTES

Conflict of interest statement

Dr. Davison reports grants from the Medical Research Council and the Wellcome Trust. Dr. Ganzenmueller reports grants from the Deutsche Forschungsgemeinschaft Collaborative Research Centre 900 and from Niedersächsische Ministerium für Wissenschaft und Kultur. Dr. Hubáček reports a grant from the Ministry of Health of the Czech Republic for the conceptual development of University Hospital, Motol, Prague, Czech Republic, personal fees and non-financial support from MSD and from Chimaerix, and personal fees from Dynex that are outside the scope of the submitted work. Dr. Lilieri reports a grant from the Fondazione Regionale per la Ricerca Biomedica, Regione Lombardia. Dr. Schulz reports grants from the Deutsche Forschungsgemeinschaft Collaborative Research Centre 900 and from the German Federal Ministry of Education and Research. Dr. Wilkinson reports a grant from the Wellcome Trust. Dr. Wilkie reports that his part in the submitted work was completed prior to his employment by Illumina.

Funding statement

This work was funded by the Medical Research Council (MC_UU_12014/3 and MC_UU_12014/12), the Wellcome Trust (204870/Z/16/Z and WT090323MA), the Ministry of Health of the Czech Republic for conceptual development of research organization 00064203 (University Hospital, Motol, Prague, Czech Republic), the Fondazione Regionale per la Ricerca Biomedica, Regione Lombardia (FRRB 2015-043), the Deutsche Forschungsgemeinschaft Collaborative Research Centre 900 (core project Z1, grant SFB-9001), the German Center of Infection Research TTU Infections of the Immunocompromised

Host, and the Niedersächsische Ministerium für Wissenschaft und Kultur (grant COALITION – Communities Allied in Infection). A. Dhingra and E. Hage were supported by the Infection Biology graduate program of Hannover Biomedical Research School.

Presentation at conferences

Parts of this work have been presented on multiple occasions, most recently at the German Society for Virology (14-17 March 2018) and the UK Microbiology Society (10-13 April 2018).

Corresponding author

Andrew J. Davison; MRC-University of Glasgow Centre for Virus Research, Sir Michael Stoker Building, 464 Bearsden Road, Glasgow G61 1QH, United Kingdom;
andrew.davison@glasgow.ac.uk; +441413306263

ABSTRACT

The genomic characteristics of human cytomegalovirus (HCMV) strains sequenced directly from clinical material were investigated, focusing on variation, multiple-strain infection, recombination and natural mutation. A total of 207 datasets generated in this and other studies using target enrichment and high-throughput sequencing were analysed, in the process facilitating the determination of genome sequences for 91 strains. Key findings were that (i) it is vital to monitor sequence data quality, especially when analysing intrahost diversity, (ii) intrahost diversity in single-strain infections is much less than that in multiple-strain infections, (iii) many recombinant strains have been generated and transmitted during HCMV evolution, and some have survived for thousands of years without further recombination, (iv) mutants lacking gene functions have been circulating and recombining for long periods and can cause congenital infection and resulting clinical sequelae. Future studies in general populations are likely to continue illuminating the evolution, epidemiology and pathogenesis of HCMV.

Keywords: human cytomegalovirus, genome sequence, target enrichment, genotype, hypervariation, multiple-strain infection, recombination, mutation, intrahost variation

BACKGROUND

Human cytomegalovirus (HCMV) poses a risk to people with immature or compromised immune systems, and can have serious outcomes in unborn children, transplant recipients and people with HIV/AIDS. Prior to the advent of high-throughput technologies, studies of HCMV genomes in natural infections were limited to Sanger sequencing of PCR products, often focusing on a small number of polymorphic (hypervariable) genes [1]. This not only ignored most of the genome, but also made it difficult to identify and characterise multiple-strain infections, which may have more serious outcomes than single-strain infections.

The first complete HCMV genome to be sequenced was that of the high-passage strain AD169, by Sanger sequencing of a set of plasmids [2]. It was over a decade before additional genomes were sequenced, also by Sanger technology, in the form of bacterial artificial chromosomes [3-5], virion DNA preparations [6] and PCR amplicons [7, 8]. These sequences were complemented by many others, most determined by high-throughput methods [7, 9-13].

With three exceptions [7, 11], all of these sequences were derived from strains that had been isolated in cell culture. Mounting data on the existence of multiple-strain infections and the propensity of HCMV to mutate during isolation [6, 7, 8, 14, 15] added impetus to sequencing genomes directly from clinical material. One strategy involves sequencing overlapping PCR amplicons [7, 16]. An alternative involves generating random DNA fragments from clinical material, amplifying them by PCR, and hybridising them to an oligonucleotide bait library representing known HCMV diversity. This target enrichment technology originated in commercial kits for cellular exome sequencing, and has been applied to various pathogens [17, 18], including HCMV [19-21]. We have used it since 2012,

and have released many genome sequences via GenBank that have been pivotal in other studies [11, 12, 19-21].

High-throughput sequencing has highlighted several features of the HCMV genome that had been discovered earlier, including variation and hypervariation [22, 25], multiple-strain infection [23], recombination [24, 25], and gene loss by natural mutation [26]. HCMV is the most variable of the human herpesviruses [12], and hypervariation of a subset of genes exists in the form of constrained genotypes that may be used to explore genome form and function. We sought to extend this process on a whole-genome basis to HCMV strains as they exist in clinical situations rather than in the laboratory.

METHODS

Samples

For convenience, samples were analysed as three collections, which are detailed in **Supplementary Tables 1-3** and summarised in **Table 1**. Collection 1 originated from congenital infections from Pavia, Jerusalem and Prague. Collection 2 originated from Hannover and Pavia, and most came from transplant recipients. Collection 3 represents samples obtained by others in previous studies from people with various conditions, and were sequenced in those studies using the approach employed here, although with a different oligonucleotide bait library. The preliminary features of the samples and datasets are in **Supplementary Tables 1-3** rows 3-7, and the clinical outcomes of congenital infections are in **Supplementary Table 1** row 207.

DNA sequencing

Target capture and library preparation were performed using the SureSelect^{XT} v. 1.7 target enrichment system for Illumina paired-end sequencing libraries with biotinylated cRNA probe baits (Agilent, Stockport, UK) [21]. Custom bait libraries representing known HCMV diversity were designed in February 2012 and April 2014 from 31 and 64 complete genome sequences, respectively. Access to the latter library is available from the corresponding author. Data on viral loads and library construction are in **Supplementary Tables 1-3** rows 9-12.

Datasets of 300 or 150 nt paired-end reads were were generated using a MiSeq instrument (Illumina, San Diego, CA, USA), and prepared for analysis using Trim Galore! v 0.4.0 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/; length=21,

quality=10 and stringency=3). The numbers of trimmed reads are in **Supplementary Tables 1-3** row 15.

Library diversity

For each dataset, the number of reads derived from unique HCMV fragments was estimated by using Bowtie2 v. 2.2.6 [29] to align the reads against the strain Merlin sequence (GenBank accession AY446894), and also, where it could be determined, the consensus genome sequence derived from the dataset (see **Table 1** for details). The relevant data are in **Supplementary Tables 1-3** rows 17-19 and 24-27. Reads containing insertions or deletions were removed from the SAM alignment file, and duplicate read pairs sharing both end coordinates, or duplicate unpaired reads sharing one end coordinate, were removed, producing an alignment file for unique reads derived from unique HCMV fragments (https://centre-for-virus-research.github.io/VATK/AssemblyPostProcessing). This file was viewed using Tablet v. 1.14.11.7 [30]. The final and initial coverage depth values, and their ratio (expressed as a percentage), are in **Supplementary Tables 1-3** rows 20-22 and 28-30. The ratio ranged from 0 to 100, with higher values indicating more diverse libraries derived from greater numbers of unique HCMV fragments.

Strain enumeration

The numbers and proportions of strains represented in each dataset were estimated by two strategies: genotype read-matching and motif read-matching (https://centre-for-virus-research.github.io/VATK/HCMV_pipeline). Both utilised datasets concatenated from the paired-end datasets. The genotype designations used were either based on reported phylogenies [6, 25, 31-33], amending or extending them as appropriate, or were onstructed

afresh using Clustal Omega v. 1.2.4 [34] and MEGA v. 6.0.6 [35] with data for the genomes listed in **Supplementary Table 4** and individual genes for which additional sequences were available in GenBank. Alignments are in **Supplementary Figure 1**.

For genotype read-matching, Bowtie2 was used to align the reads to sequences representing the genotypes of two hypervariable genes, UL146 and RL13 [6, 12, 36]. The sequences are in **Supplementary Tables 1-3** rows 36-60, and represent the entire coding region of UL146 and the central coding region of RL13. In contrast to the UL146 genotypes, the RL13 genotypes cross-matched to some extent in four groups (G1, G2, G3; G4A, G4B; G6, G10; and G7, G8). In these instances, the genotype within the group with most matching reads was scored. The numbers of reads aligned to each genotype are also in **Supplementary Tables 1-3** rows 36-60. A genotype was scored if the number of reads was >10 and also represented >2% of the total number detected for all genotypes of that gene. For 14 samples in collection 1 that had been sequenced prior to the availability of ultrapure (TruGrade) oligonucleotides, these values were set at >25 and >5%, respectively. The number of strains in a sample was scored as the greater of the numbers of genotypes detected for the two target genes, and is in **Supplementary Tables 1-3** row 13.

For motif read-matching, conserved genotype-specific motifs (20-31 nt) were identified manually for 12 hypervariable genes [6, 12, 19, 33]. Additional motifs were included in order to identify certain recombinants or mutants. The motif sequences and number of reads containing perfect matches to the motif or its reverse complement are in **Supplementary Tables 1-3** rows 62-180. Genotypes were scored as described above. The number of strains in a sample was scored at the greatest of the numbers of genotypes detected among the target genes, with a requirement that at least this number should have been detected for at least two genes, and is in **Supplementary Tables 1-3** row 14.

202

203 **Data deposition**

204 The original read datasets (purged of human data) were deposited in ENA (project no.
205 PRJEB29585), and the consensus genome sequences were deposited in GenBank, under the
206 accession numbers in **Supplementary Tables 1-3** rows 8 and 31, respectively. Updated
207 genome sequences in collection 3 were deposited by the original submitters in GenBank
208 [19] or by us as third-party annotations in ENA (project no. PRJEB29374) [20]. Features of
209 the sequences are in **Supplementary Tables 1-3** rows 32-34. Data on mutants in collection 1
210 are in **Supplementary Table 1** rows 182-205.

211

212 **Intrahost variation**

213 Variation was examined in datasets for which a consensus genome sequence had been
214 determined. Original datasets were prepared for analysis using Trim Galore! (length=100,
215 quality=30 and stringency=1), and trimmed reads were mapped using Bowtie2. Alignment
216 files in SAM format were converted into BAM format, sorted using SAMtools v. 1.3 [37], and
217 analysed using LoFreq v. 2.1.2 [38] and V-Phaser 2 [39] under default parameters.

RESULTS

Operational limitations

The analysis involved a total of 207 datasets from 199 samples and 102 individuals (**Table 1** and **Supplementary Tables 1-3**). The percentage of HCMV reads (target enrichment efficiency) and the coverage ratio (library diversity) tended to depend on the sample type (proportion of host DNA) and the number of genome copies used to make the library. In general, >1000 copies per library were needed to obtain data of sufficient quality to determine a complete genome sequence. However, data quality was influenced by many factors, including logistical errors, low coverage depth, low library diversity, and the apparent presence of additional strains at levels below the threshold, as a result of low-level multiple-strain infection or cross-contamination. In addition, an inability to obtain data from the entire genome, despite excellent coverage ratios, precluded determination of complete sequences from most datasets in collection 3.

Genome sequences

A total of 91 complete or almost complete HCMV genome sequences were determined (**Table 1**). We reported five of these previously [21], and 16 are improvements on published sequences [19]. Most originated from single-strain infections or multiple-strain infections in which one strain was predominant, and some originated from different strains that predominated in a patient at different times. Defining a strain as a virus present in an individual, these 91 sequences, plus an additional 49 deposited by our group and 104 by other groups, brought the number of strains sequenced to 244 (**Supplementary Table 4**). Of these, 91 were sequenced directly from clinical material, and all but one of these were

242 determined by us. The mean size of the HCMV genome, based on the 219 complete
243 sequences lacking sizeable deletions, is 235,514 bp.
244
245 **Multiple-strain infections**
246 Genotypic differences in hypervariable genes (**Figure 1** and **Supplementary Figure 1**)
247 were exploited to detect multiple-strain infections by genotype read-matching and motif
248 read-matching , the latter proving to be the more versatile method. Single strains were
249 present in the great majority of congenitally infected patients (n=43/50 in collections 1 and
250 2), whereas they were significantly less common in transplant recipients (n=9/23 in
251 collections 2 and 3; Pearson's chi-squared test, χ^2=14.678, p<0.05).
252
253 **Recombination**
254 The 244 genome sequences were genotyped in the 12 hypervariable genes used for
255 motif read-matching and then in five more (**Figure 1** and **Supplementary Table 4**).
256 Hypervariation in UL55, which encodes glycoprotein B (gB), is located in two regions
257 (UL55N, near the N terminus, and UL55X, encompassing the proteolytic cleavage site) [23,
258 40]. Five genotypes (G1-G5) have been assigned to each region [23, 40-42], which are
259 separated by 927 bp that are 80% identical in all strains. All genomes had one of the
260 previously reported UL55X genotypes (**Supplementary Table 5**). As reported previously [40],
261 UL55N G2 and G3 could not be distinguished reliably from each other, and two additional
262 genotypes (G6-G7) were detected that may have arisen from ancient recombination events
263 within UL55N (**Supplementary Tables 4 and 5**, and **Supplementary Figure 1**). There was
264 evidence for recombination in the region between UL55N and UL55X in only eight genomes.
265 This low proportion of recombination (3.3%) in a small region contrasts with higher levels

proposed proposed previously [40, 43], which may have been influenced by PCR-based artefacts arising from the presence of multiple strains. UL73 and UL74, which encode glycoproteins N and O, respectively, are adjacent hypervariable genes that exist as eight genotypes each [25, 32, 44]. There was evidence for recombination between them in only seven genomes (2.9%), in accordance with the low levels (4%) detected in PCR-based studies [25, 32]. In the region containing adjacent hypervariable genes RL12, RL13 and UL1, recombinants were rare (1.2%) in RL12 and absent from RL13 and UL1. In contrast, hypervariable genes UL146 and UL139, which encode a CXC chemokine and a membrane glycoprotein, respectively, are separated by a well-conserved region of over 5 kbp. The number (66) of the 126 possible genotype combinations represented in the 244 genomes is too large to allow any underlying genotypic linkage to be discerned, consistent with previous conclusions based on PCR [45]. No recombinants were noted within UL146.

In principle, strains in multiple-strain infections have the opportunity to recombine. In our previous analysis of RTR1 in collection 2, we noted that one strain predominated at earlier times and another at later times [21]. From the low frequency of variants across a large part of the genome, we concluded that the second strain had arisen either by recombination involving the first strain or by reinfection with, or reactivation of, a second strain fortuitously similar to the first. Here, recombination was strongly supported by a comparison of the two genome sequences (RTR1A and RTR1B), which showed that approximately two-thirds of the genome is almost identical (differing by three substitutions in noncoding regions), whereas the remaining third of the genome is highly dissimilar.

To investigate whether strains have been transmitted without recombination occurring, identical genotypic constellations were identified among the 244 genomes (**Supplementary Table 6**). This revealed the existence of 12 haplotype groups within which multiple strains

exhibit no signs of having recombined since diverging from their last common ancestor; these are termed nonrecombinant strains below. These results suggest that nonrecombinant strains have been circulating, some for periods sufficient to allow the accumulation of >100 substitutions. Application of an evolutionary rate estimated for herpesviruses (3.5×10^{-8} substitutions/nt/year) [46] implies that these periods may have extended to many thousands of years. The distribution of substitutions across the genome in highly divergent groups 9 and 10 was examined in further detail. Group 9 (three strains) exhibited 135 differences, with the 50 that would affect protein coding distributed among 38 genes, and group 10 (two strains) exhibited 138 differences, with the 38 that would affect protein coding distributed among 27 genes. No obvious bias was observed towards greater diversity in any particular gene or group of genes, including those in the hypervariable category. As suggested by the lack of diversity within genotypes in comparison with the marked diversity among them (**Supplementary Figure 1**), these results fit with the view that intense diversification of hypervariable genes occurred early in human or pre-human history [30, 45], and has long since ceased.

Mutations

Mutations that cause premature translational termination, and therefore potentially affect gene function, have been catalogued previously in HCMV genomes [7, 11, 12, 26]. They may have resulted from substitutions that introduce in-frame stop codons or affect the conserved dinucleotide at the beginning or end of an intron, or from insertions, deletions or inversions that cause frameshifting or loss of protein-coding regions. The underlying data have been derived mostly from strains isolated in cell culture, and their interpretation has assumed that these mutations occurred naturally. For example, in a major study reporting

that 75% of strains are mutated [12], 157 mutations were identified in 101 strains (100 passaged *in vitro*), but only 35 were confirmed in the clinical material. Nonetheless, the distribution of mutations among the 91 strains sequenced directly from clinical material appears similar to that among passaged strains (**Table 2** and **Supplementary Table 4**).

Among the strains sequenced from clinical material, 77% are mutated in at least one gene (compared with 79% among all sequenced strains), and one is mutated in as many as six genes (Pat_D in collection 3). In clinical strains, the most frequently mutated genes are UL9, RL5A, UL1 and RL6 (members of the RL11 family), US7 and US9 (members of the US6 gene family), and UL111A (encoding viral interleukin-10) (**Supplementary Figure 1**). The likelihood that many of these mutations were not generated in the patients concerned but are ancient is supported by the finding that all were detected at levels very close to 100% in collection 1, and by previous observations that the same mutation is present in different strains [7, 12]. In addition, there was evidence from the PAV6 datasets for maternal transmission of a US7 mutant (**Supplementary Table 1**), and from PCR data (not shown) of a UL111A mutant to PAV16. Moreover, nine of the groups of nonrecombinant strains contained mutants, and some of the mutations were common to group members (**Supplementary Table 6**) and even to additional strains among the 244, indicating that they had been transferred by recombination. These observations again indicate the longevity of many mutations and their propagation by recombination. Focusing on the most common mutations, strains in which UL9, RL5A, UL1, US9, US7 and UL111A were affected (singly or in combination) were, like nonmutated strains, associated with congenital infection and, in some cases, defects in neurological development (**Supplementary Table 1**).

Intrahost diversity

Use of LoFreq and V-Phaser showed that single-strain infections contained markedly fewer variants (median values of 60 and 140, respectively) than multiple-strain infections (median values of 2444 and 2955, respectively; **Figure 2**). The differences between the values for single- and multiple-strain infections were assessed as being significant by the Kruskal-Wallis rank-sum test (LoFreq, Kruskal-Wallis χ^2=67.918, p<2.2e^{-16}; V-Phaser, Kruskal-Wallis χ^2=63.536, p=1.6e^{-15}). Seven outliers in single-strain infections were common to both analyses (in order of decreasing number of variants, RTR6B, CMV-37, RTR2, CMV-35, CMV-38, ERR1279054 and PAV6), one was reported by LoFreq only (PAV21), and four were reported by V-Phaser only (CMV-19, CMV-31, PRA6A and SCRT12).

DISCUSSION

Advances in high-throughput sequencing technology have made it possible to generate a wealth of viral genome information directly from clinical material. However, operational factors should be taken into account in assessing the data. These include sample characteristics (source, viral content and presence of multiple strains), confounding events (logistical errors and contamination), adequate design of the bait library and the sequencing protocol (ability to enrich all variants and acquire data evenly across the genome), and quality and extent of sequencing data (library diversity and coverage depth). Since perceived levels of intrahost variation are particularly sensitive to these factors, and may be greatly over- or underestimated as a result, we proceeded cautiously. However, as indicated in our earlier study [21], it is clear that the number of variants in single-strain infections was markedly less than that in multiple-strain infections. Moreover, it was also far less than that reported by others in samples from congenital infections [16]. The factors listed above may have been responsible for the outliers in single-strain infections, and may also have generally influenced the derived median values. In our view, accurate estimates of intrahost variation in single-strain infections are not yet available, and will require sequencing and bioinformatic approaches that are demonstrably robust, reliable and reproducible [47, 48].

Whole-genome analyses have confirmed the significant role of recombination during HCMV evolution reported in numerous earlier studies [12, 19]. Recombination has occurred over a very long period and remains limited in extent, with surviving recombination events being more common in long regions, less common in short regions, and rare in hypervariable regions, as would be consistent with homologous recombination. It is possible that recombination frequency is restricted in some circumstances by functional

interdependence of regions in the same protein (e.g. gB) or separate proteins (e.g. gN and gO [30, 44]), but it is not known whether differential recombination due to sequence relatedness is of general biological significance. Also, strains have circulated that seem not to have recombined over many thousands of years. The extent to which recombinants arise and survive in individuals with multiple-strain infections is a different question, particularly in immunosuppressed transplantation patients. Except where populations fluctuate significantly and are sampled serially (e.g. RTR1 in collection 2), this is difficult to approach using short-read data, as these are based on PCR methodologies prone to generating recombinational artefacts. Long- or single-read technologies and new bioinformatic tools should contribute significantly to this area. Also, conclusions drawn from transplant recipients, who are immunosuppressed and in whom HCMV populations may be diversified by transplantation from HCMV-positive donors or selected with antiviral drugs, are unlikely to represent natural situations. More relevant are maternal transmission routes, including those involving breast milk (Suárez *et al.*, manuscript in preparation).

The frequent identification of mutants, and their apparently long history, reveals an interesting aspect of HCMV microevolution. The implication that some mutants have a selective advantage in certain circumstances may be extended to their evident ability nonetheless to cause congenital infections and associated neurological sequelae, probably in combination with specific host factors. Mutated genes tend to be involved, or are suspected to be involved, in immune modulation. These genes include UL111A, which encodes viral interleukin-10 [49], and UL40, which is involved in protecting infected cells against NK cell lysis [50] via its signal peptide, in which mutations in this gene occur. By analogy to other members of the RL11 family, the most frequently mutated gene (UL9), is also likely to be involved in an aspect of immune modulation.

Modern approaches offer a powerful means for analysing HCMV genomes directly from clinical material, with the important proviso that the data should be monitored for quality and interpreted in the context of the known evolutionary and biological characteristics of the virus. The sequence data promise to become very extensive and will help shed further light on the epidemiology, pathogenesis and evolution of HCMV in clinical and natural setting, thus allowing investigation of virulence determinants and the development of new interventions.

ACKNOWLEDGMENTS

We are very grateful Florent Lasalle, Daniel Depledge and Judith Breuer (University College London) for kindly providing unpublished collection 3 datasets and for updating the associated genome sequences in GenBank. We also thank Jenny Witthuhn (Hannover Medical School) for excellent technical assistance.

REFERENCES

1. Puchhammer-Stöckl E, Görzer I. Cytomegalovirus and Epstein-Barr virus subtypes – the search for clinical significance. J Clin Virol 2006; 36:239-48.
2. Chee MS, Bankier AT, Beck S, et al. Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. Curr Top Microbiol Immunol 1990; 154:125-69.
3. Dunn W, Chou C, Li H, et al. Functional profiling of a human cytomegalovirus genome. Proc Natl Acad Sci U S A 2003; 100:14223-8.
4. Murphy E, Yu D, Grimwood J, et al. Coding potential of laboratory and clinical strains of human cytomegalovirus. Proc Natl Acad Sci U S A 2003; 100:14976-81.
5. Sinzger C, Hahn G, Digel M, et al. Cloning and sequencing of a highly productive, endotheliotropic virus strain derived from human cytomegalovirus TB40/E. J Gen Virol 2008; 89:359-68.
6. Dolan A, Cunningham C, Hector RD, et al. Genetic content of wild-type human cytomegalovirus. J Gen Virol 2004; 85:1301-12.
7. Cunningham C, Gatherer D, Hilfrich B, et al. Sequences of complete human cytomegalovirus genomes from infected cell cultures and clinical specimens. J Gen Virol 2010; 91:605-15.
8. Dargan DJ, Douglas E, Cunningham C, et al. Sequential mutations associated with adaptation of human cytomegalovirus to growth in cell culture. J Gen Virol 2010; 91:1535-46.
9. Bradley AJ, Lurain NS, Ghazal P, et al. High-throughput sequence analysis of variants of human cytomegalovirus strains Towne and AD169. J Gen Virol 2009; 90:2375-80.

10. Jung GS, Kim YY, Kim JI, et al. Full genome sequencing and analysis of human cytomegalovirus strain JHC isolated from a Korean patient. Virus Res 2011; 156:113-20.
11. Sijmons S, Thys K, Corthout M, et al. A method enabling high-throughput sequencing of human cytomegalovirus complete genomes from clinical isolates. PLoS One 2014; 9:e95501.
12. Sijmons S, Thys K, Mbong Ngwese M, et al. High-throughput analysis of human cytomegalovirus genome diversity highlights the widespread occurrence of gene-disrupting mutations and pervasive recombination. J Virol 2015; 89:7673-95.
13. Zhao F, Shen ZZ, Liu ZY, et al. Identification and BAC construction of Han, the first characterized HCMV clinical strain in China. J Med Virol 2016; 88:859-70.
14. Cha TA, Tom E, Kemble GW, Duke GM, Mocarski ES, Spaete RR. Human cytomegalovirus clinical isolates carry at least 19 genes not found in laboratory strains. J Virol 1996; 70:78-83.
15. Stanton RJ, Baluchova K, Dargan DJ, et al. Reconstruction of the complete human cytomegalovirus genome in a BAC reveals RL13 to be a potent inhibitor of replication. J Clin Invest 2010; 120:3191-208.
16. Renzette N, Bhattacharjee B, Jensen JD, Gibson L, Kowalik TF. Extensive genome-wide variability of human cytomegalovirus in congenitally infected infants. PLoS Pathog 2011; 7:e1001344.
17. Melnikov A, Galinsky K, Rogov P, et al. Hybrid selection for sequencing pathogen genomes from clinical samples. Genome Biol 2011; 12:R73.
18. Depledge DP, Palser AL, Watson SJ, et al. Specific capture and whole-genome sequencing of viruses from clinical samples. PLoS One 2011; 6:e27805.

454 19. Lassalle F, Depledge DP, Reeves MB, et al. Islands of linkage in an ocean of pervasive
455 recombination reveals two-speed evolution of human cytomegalovirus genomes. Virus
456 Evol 2016; 2:vew017.
457 20. Houldcroft CJ, Bryant JM, Depledge DP, et al. Detection of low frequency multi-drug
458 resistance and novel putative Maribavir resistance in immunocompromised pediatric
459 patients with cytomegalovirus. Front Microbiol 2016; 7:1317.
460 21. Hage E, Wilkie GS, Linnenweber-Held S, et al. Characterization of human
461 cytomegalovirus genome diversity in immunocompromised hosts by whole-genome
462 sequencing directly from clinical specimens. J Infect Dis 2017; 215:1673-83.
463 22. Chou SW, Dennison KM. Analysis of interstrain variation in cytomegalovirus glycoprotein
464 B sequences encoding neutralization-related epitopes. J Infect Dis 1991; 163:1229-34.
465 23. Meyer-König U, Ebert K, Schrage B, Pollak S, Hufert FT. Simultaneous infection of healthy
466 people with multiple human cytomegalovirus strains. Lancet 1998; 352:1280-1.
467 24. Rasmussen L, Geissler A, Winters M. Inter- and intragenic variations complicate the
468 molecular epidemiology of human cytomegalovirus. J Infect Dis 2003; 187:809-19.
469 25. Mattick C, Dewin D, Polley S, et al. Linkage of human cytomegalovirus glycoprotein gO
470 variant groups identified from worldwide clinical isolates with gN genotypes,
471 implications for disease associations and evidence for N-terminal sites of positive
472 selection. Virology 2004; 318:582-97.
473 26. Sekulin K, Görzer I, Heiss-Czedik D, Puchhammer-Stöckl E. Analysis of the variability of
474 CMV strains in the RL11D domain of the RL11 multigene family. Virus Genes 2007;
475 35:577-83.
476 27. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its
477 applications to single-cell sequencing. J Comput Biol 2012; 19:455-77.

478 28. Silva GG, Dutilh BE, Matthews TD, et al. Combining de novo and reference-guided
479 assembly with scaffold_builder. Source Code Biol Med. 2013; 8:23.
480 29. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;
481 9:357-9.
482 30. Milne I, Stephen G, Bayer M, et al. Using Tablet for visual exploration of second-
483 generation sequencing data. Brief Bioinform 2013; 14:193-202.
484 31. Bradley AJ, Kovács IJ, Gatherer D, et al. Genotypic analysis of two hypervariable human
485 cytomegalovirus genes. J Med Virol 2008; 80:1615-23.
486 32. Bates M, Monze M, Bima H, Kapambwe M, Kasolo FC, Gompels UA; CIGNIS study group.
487 High human cytomegalovirus loads and diverse linked variable genotypes in both HIV-1
488 infected and exposed, but uninfected, children in Africa. Virology 2008; 382:28-36.
489 33. Sijmons S. Diversity and evolution of human cytomegalovirus. Katholieke Universiteit
490 Leuven. Leuven, Belgium. 2015. (PhD thesis).
491 34. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein
492 multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011; 7:539.
493 35. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary
494 Genetics Analysis version 6.0. Mol Biol Evol 2013; 30:2725-9.
495 36. Davison AJ, Holton M, Dolan A, Dargan DJ, Gatherer D, Hayward GS. Comparative
496 genomics of primate cytomegaloviruses. In: Reddehase MJ, ed. Cytomegaloviruses: from
497 molecular pathogenesis to intervention. Vol. 1. Norwich, UK: Caister Academic Press,
498 2013.
499 37. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and
500 SAMtools. Bioinformatics 2009; 25:2078-9.

501 38. Wilm A, Aw PP, Bertrand D, et al. LoFreq: a sequence-quality aware, ultra-sensitive
502 variant caller for uncovering cell-population heterogeneity from high-throughput
503 sequencing datasets. Nucleic Acids Res 2012; 40:11189-201.
504 39. Yang X, Charlebois P, Macalalad A, Henn MR, Zody MC. V-Phaser 2: variant inference for
505 viral populations. BMC Genomics 2013; 14:674.
506 40. Meyer-König U, Haberland M, von Laer D, Haller O, Hufert FT. Intragenic variability of
507 human cytomegalovirus glycoprotein B in clinical strains. J Infect Dis 1998; 177: 1162-9.
508 41. Shepp DH, Match ME, Lipson SM, Pergolizzi RG. A fifth human cytomegalovirus
509 glycoprotein B genotype. Res Virol 1998; 149:109-14.
510 42. Deckers M, Hofmann J, Kreuzer KA, et al. High genotypic diversity and a novel variant of
511 human cytomegalovirus revealed by combined UL33/UL55 genotyping with broad-range
512 PCR. Virol J 2009; 6:210.
513 43. Haberland M, Meyer-König U, Hufert FT. Variation within the glycoprotein B gene of
514 human cytomegalovirus is due to homologous recombination. J Gen Virol 1999;
515 80:1495-500.
516 44. Paterson DA, Dyer AP, Milne RS, Sevilla-Reyes E, Gompels UA. A role for human
517 cytomegalovirus glycoprotein O (gO) in cell fusion and a new hypervariable locus.
518 Virology 2002; 293:281-94.
519 45. Bradley AJ, Kovács IJ, Gatherer D, et al. Genotypic analysis of two hypervariable human
520 cytomegalovirus genes. J Med Virol 2008; 80:1615-23.
521 46. McGeoch DJ, Cook S, Dolan A, Jamieson FE, Telford EA. Molecular phylogeny and
522 evolutionary timescale for the family of mammalian herpesviruses. J Mol Biol 1995;
523 247:443-58.

524 47. Xu C, Nezami Ranjbar MR, Wu Z, DiCarlo J, Wang Y. Detecting very low allele fraction
525 variants using targeted DNA sequencing and a novel molecular barcode-aware variant
526 caller. BMC Genomics 2017; 18:5.
527 48. Illingworth CJR, Roy S, Beale MA, Tutill H, Williams R, Breuer J. On the effective depth of
528 viral sequence data. Virus Evol 2017; 3:vex030.
529 49. McSharry BP, Avdic S, Slobedman B. Human cytomegalovirus encoded homologs of
530 cytokines, chemokines and their receptors: roles in immunomodulation. Viruses 2012;
531 4:2448-70.
532 50. Prod'homme V, Tomasec P, Cunningham C, et al. Human cytomegalovirus UL40 signal
533 peptide regulates cell surface expression of the Natural Killer cell ligands HLA-E and
534 gpUL18. J Immunol 2012; 188:2794-804.

535 **FIGURE LEGENDS**

536

537 **Figure 1**

538 Locations in the HCMV strain Merlin genome of genes used for genotyping. The genome
539 consists of two unique regions, U_L (1325-194343 bp) and U_S (197627-233108 bp), the former
540 flanked by inverted repeats TR_L (1-1324 bp) and IR_L (194344-195667 bp), and the latter
541 flanked by inverted repeats IR_S (195090-197626 bp) and TR_S (233109- 235646 bp). Protein-
542 coding regions are indicated by colour-shaded arrows, and noncoding RNAs as narrower,
543 white-shaded arrows, with gene nomenclature below. Introns are shown as narrow white
544 bars. The 12 genes (RL5A, RL6, RL12, RL13, UL1, UL9, UL11, UL73, UL74, UL120, UL146 and
545 UL139) used for motif read-matching (**Supplementary Tables 1-3**) are coloured red. Two of
546 these genes (RL13 and UL146) were also used for genotype read-matching (**Supplementary**
547 **Tables 1-3**). The additional five genes (UL20, UL33, UL37, UL55 and US9) used to genotype
548 genome sequences by alignment (**Supplementary Table 4**) are coloured orange.

549

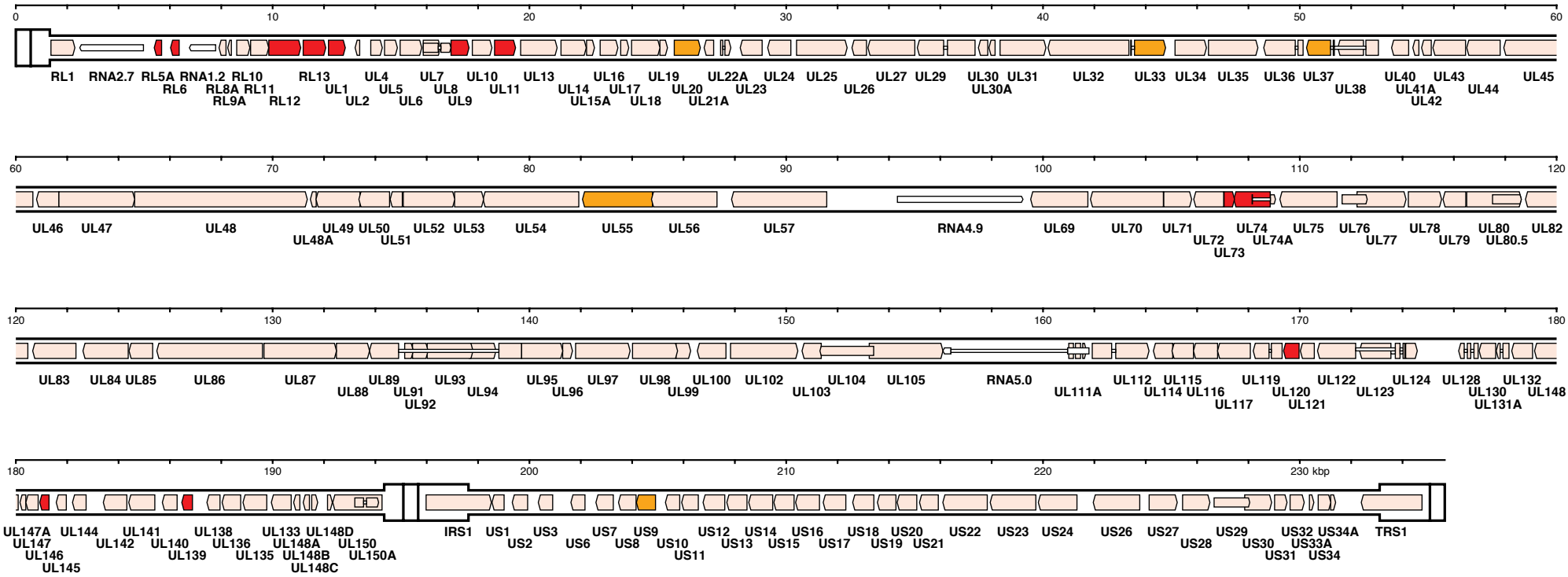
550 **Figure 2**

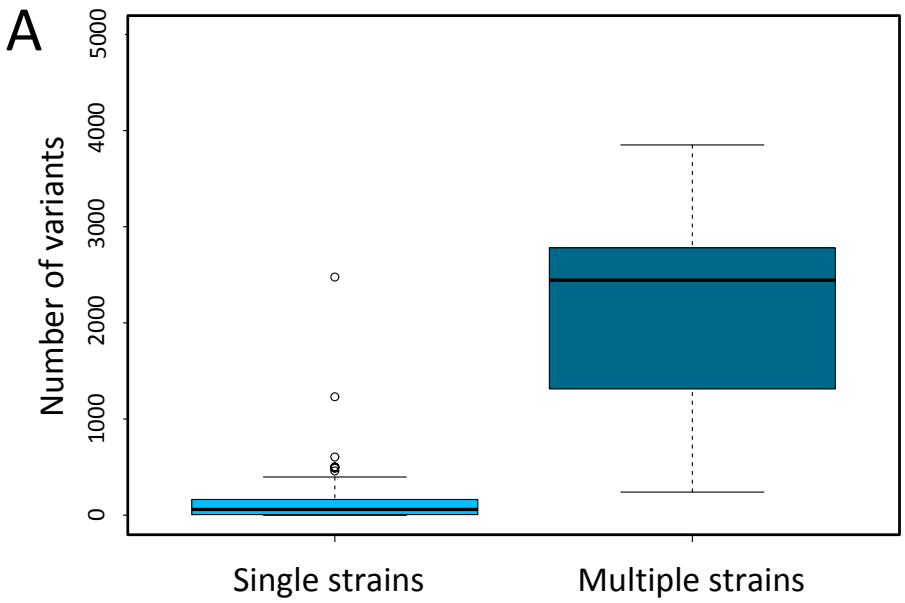
551 Box-and-whisker graphs showing the number of variants detected in single-strain and
552 multiple-strain infections using (A) LoFreq and (B) V-Phaser. Single-strain (n=134 and 131,
553 respectively) and multiple-strain datasets (n=29 and 29, respectively) for which consensus
554 genome sequences were available were identified by motif read-matching (see Methods;
555 **Supplementary Tables 1-3**). The total number of variants (2-50% of the total) in each
556 dataset was enumerated; length polymorphisms were not considered. Each box
557 encompasses the first to third quartiles (Q1 to Q3) and shows the median as a thick line. For
558 each box, the horizontal line at the end of the upper dashed whisker marks the upper

559 extreme (defined as the smaller of Q3+1.5(Q3-Q1) and the highest single value), and the

560 horizontal line at the end of the lower dashed whisker marks marks the lower extreme (the

561 greater of Q1-1.5(Q3-Q1) and the lowest single value).


0
10
20
30
40
50
60
RL1 RNA2.7 RL5A RL6 RNA1.2 RL8A RL9A RL10 RL11 RL12 RL13 UL1 UL2 UL4 UL5 UL6 UL7 UL8 UL9 UL10 UL11 UL13 UL14 UL15A UL16 UL17 UL18 UL19 UL20 UL21A UL22A UL23 UL24 UL25 UL26 UL27 UL29 UL30 UL30A UL31 UL32 UL33 UL34 UL35 UL36 UL37 UL38 UL40 UL41A UL42 UL43 UL44 UL45
60
70
80
90
100
110
120
UL46 UL47 UL48 UL48A UL49 UL50 UL51 UL52 UL53 UL54 UL55 UL56 UL57 RNA4.9 UL69 UL70 UL71 UL72 UL73 UL74 UL74A UL75 UL76 UL77 UL78 UL79 UL80 UL80.5 UL82
120
130
140
150
160
170
180
UL83 UL84 UL85 UL86 UL87 UL88 UL89 UL91 UL92 UL93 UL94 UL95 UL96 UL97 UL98 UL99 UL100 UL102 UL103 UL104 UL105 RNA5.0 UL111A UL112 UL114 UL115 UL116 UL117 UL119 UL120 UL121 UL122 UL123 UL124 UL128 UL130 UL131A UL132 UL148
180
190
200
210
220
230 kbp
UL147A UL147 UL146 UL145 UL144 UL142 UL141 UL140 UL139 UL138 UL136 UL135 UL133 UL148A UL148B UL148C UL148D UL150 UL150A IRS1 US1 US2 US3 US6 US7 US8 US9 US10 US11 US12 US13 US14 US15 US16 US17 US18 US19 US20 US21 US22 US23 US24 US26 US27 US28 US29 US30 US31 US32 US33A US34 US34A TRS1



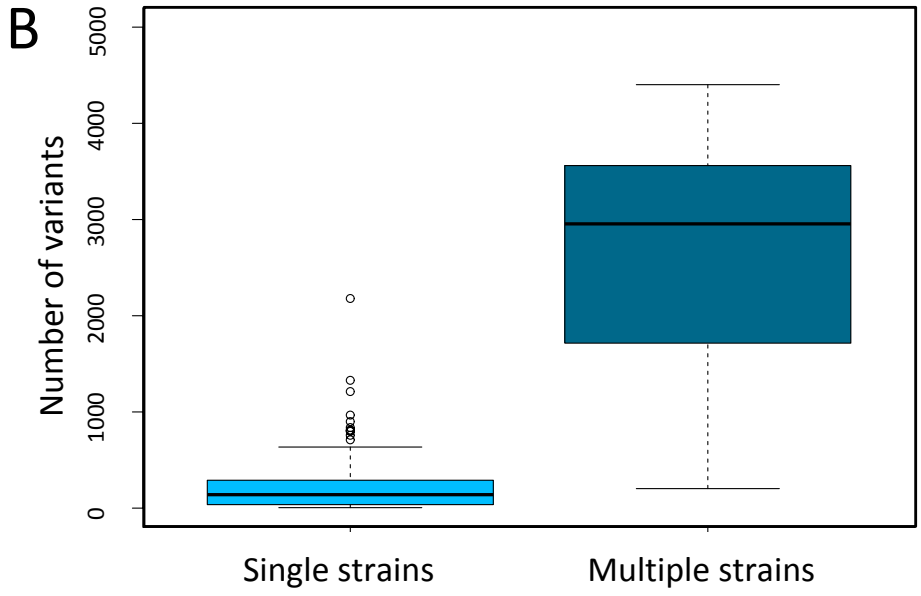


Table 1. Selected information on the sample collections. Full details are provided in Supplementary Tables 1-3.

Collection	1	2	3
Patients (no.)[a]	48	29	25
Samples (no.)	53	89	57
Sample source and prefix	Pavia (PAV), Jerusalem (JER), Prague (PRA)	Hannover (Child, RTR, SCTR), Pavia (PAV)	Rotterdam (Rot), London (Lon, Pat_)
Datasets (no.)	53	97[b]	57[c]
Duplicated libraries (no.)	0	7	0
HCMV load (IU/µl)[d]	26-559,968	5-194,840	104-18,377
Genome copies for library (no.)[e]	225-8,399,520	280-3,896,800	unknown
Reads in Merlin alignment (%)	2-91	0-85	0-90
Coverage ratio in Merlin alignment (% unique/total reads)	0.40-83.12	0.00-76.09	0.00-90.21
Genome sequences determined (no.)[f]	42	25	24

[a]Archived diagnostic samples were used, and clinical data were retrieved from clinical records, with the approval of the institutional review boards of Policlinico San Matteo, Pavia (reference nos. 35853/2010 and 35854/2010), Hadassah University Hospital, Jerusalem (reference no. HMO-063911), Motol University Hospital, Prague (reference no. EK-701a/16) and Hannover Medical School, Hannover (reference no. 2527-2014).

[b]We reported 68 of the Hannover datasets previously [21].

[c]These datasets were reported previously by others, and were either provided by the authors [19] or downloaded from ENA (study PRJEB12814) [20].

[d]HCMV load in most extracted samples was quantified in the laboratory of origin or in the sequencing laboratory. In some instances, the entire sample was used blind to generate a sequencing library.

[e]Assumes that 1 IU is equivalent to 1 genome copy.

[f]The trimmed paired-read data were aligned to the UCSC hg19 human reference genome (http://genome.ucsc.edu/) using Bowtie2. Nonmatching reads were assembled *de novo* into contigs using SPAdes 3.5.0 [27]. The contigs were ordered using Scaffold_builder v. 2.2 [28] by reference to a version of the strain Merlin sequence lacking all but 100 nt of the terminal repeat regions (TR_L at the left end and TR_S at the right end; **Figure 1**), and merged into a draft genome sequence. Residual gaps were filled by identifying relevant reads anchored in flanking regions and assembling them manually in a reiterative fashion. TR_L and TR_S were reinstated, and the complete genome sequence was verified by aligning it against the read data using Bowtie2 and inspecting the alignment in Tablet. An annotated genome sequence was produced using Sequin (https://www.ncbi.nlm.nih.gov/Sequin/).

Table 2. Mutated genes in order of decreasing frequency.

Gene	Strains mutated (no.)[a]			Strains mutated (%)[a]		
	Passaged[b]	Clinical[c]	All[d]	Passaged[b]	Clinical[c]	All[d]
UL9	50	31	81	32.89	34.07	33.33
RL5A	31	27	58	20.39	29.67	23.87
UL1	20	18	38	13.16	19.78	15.64
RL6	23	14	37	15.13	15.38	15.23
US9	26	11	37	17.11	12.09	15.23
UL111A	16	7	23	10.53	7.69	9.47
UL150	11	3	14	7.24	3.30	5.76
US7	7	7	14	4.61	7.69	5.76
UL40	8	2	10	5.26	2.20	4.12
UL30	2	3	5	1.32	3.30	2.06
UL142	2	3	5	1.32	3.30	2.06
RL12	3	1	4	1.97	1.10	1.65
RL1	1	2	3	0.66	2.20	1.23
UL136	3	0	3	1.97	0.00	1.23
US13	3	0	3	1.97	0.00	1.23
UL133	2	0	2	1.32	0.00	0.82
US6	1	1	2	0.66	1.10	0.82
US8	0	2	2	0.00	2.20	0.82
US27	2	0	2	1.32	0.00	0.82
UL11	1	0	1	0.66	0.00	0.41
UL13	0	1	1	0.00	1.10	0.41
UL14	0	1	1	0.00	1.10	0.41
UL15A	0	1	1	0.00	1.10	0.41
UL20	1	0	1	0.66	0.00	0.41
UL43	0	1	1	0.00	1.10	0.41
UL99	1	0	1	0.66	0.00	0.41
UL148	1	0	1	0.66	0.00	0.41
UL147	1	0	1	0.66	0.00	0.41
UL145	0	1	1	0.00	1.10	0.41
UL150A	1	0	1	0.66	0.00	0.41
IRS1	1	0	1	0.66	0.00	0.41
US1	1	0	1	0.66	0.00	0.41
US12	1	0	1	0.66	0.00	0.41
US19	0	1	1	0.00	1.10	0.41

[a]Omitting mutations that occurred in RL13, UL128, UL130 and UL131A probably during passage, or that were engineered during bacterial artificial chromosome construction.
[b]Strains sequenced from strains passaged in cell culture, not taking into account the minority of mutations confirmed from the clinical samples (n=151, excludes CZ/3/2012, which is the same strain as PRA8; **Supplementary Table 6**).
[c]Strains sequenced directly from clinical material (n=91).
[d]Strains sequenced directly from clinical material or passaged virus (n=243).