Major article

Human Cytomegalovirus Genomes Sequenced Directly from Clinical Material: Variation, Multiple-Strain Infection, Recombination and Mutation

Nicolás M. Suárez[1a], Gavin S. Wilkie[1ab], Elias Hage[2,3], Salvatore Camiolo[1], Marylouisa Holton[1c], Joseph Hughes[1], Maha Maabar[1d], 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

Running title: Clinical HCMV genomes

Abstract: 150 words

Main text: 3500 words

Presented in part most recently at the German Society for Virology (March 2018, Würzburg, Germany) and the UK Microbiology Society (April 2018, Birmingham, United Kingdom).

A preprint version of this paper was published in December 2018 (BioRxiv 505735. doi: https://doi.org/10.1101/505735).

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

[b]Present affiliation: Illumina, Scoreseby, Victoria, Australia.

[c]Present affiliation: SGS Vitrology Ltd., Glasgow, United Kingdom.

[d]Present affiliation: IT Services – Business Systems Team, University of Glasgow, Glasgow, United Kingdom.

Correspondence: A. 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).

ABSTRACT

The genomic characteristics of human cytomegalovirus (HCMV) strains sequenced directly from clinical pathology samples were investigated, focusing on variation, multiple-strain infection, recombination and natural mutation. A total of 207 datasets generated in this and previous 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 important to monitor the quality of sequencing libraries in investigating diversity, (ii) many recombinant strains have been transmitted during HCMV evolution, and some have apparently survived for thousands of years without further recombination, (iii) mutants with non-functional genes (pseudogenes) have been circulating and recombining for long periods and can cause congenital infection and resulting clinical sequelae, and (iv) intrahost diversity in single-strain infections is much less than that in multiple-strain infections. Future population-based studies 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

65 **BACKGROUND**

66

67 Human cytomegalovirus (HCMV) poses a risk particularly to people with immature or
68 compromised immune systems, and can have serious outcomes in congenitally infected
69 children, transplant recipients and people affected by HIV/AIDS. Prior to the advent of high-
70 throughput technologies, studies of HCMV genomes in natural infections were limited to
71 Sanger sequencing of PCR amplicons, often focusing on a small number of polymorphic
72 (hypervariable) genes [1]. This left out most of the genome and also restricted the
73 characterisation of multiple-strain infections, which may have more serious outcomes.
74 The first complete HCMV genome sequence to be determined was that of the high-
75 passage strain AD169 [2], from a plasmid library. Over a decade later, additional genomes
76 were sequenced from bacterial artificial chromosomes [3-5], virion DNA [6] or overlapping
77 PCR amplicons [7, 8]. These sequences were determined using Sanger technology, and were
78 complemented subsequently by many others, increasingly using high-throughput methods
79 [7, 9-13]. With only three exceptions [7, 11], all were derived from laboratory strains
80 isolated in cell culture. Mounting evidence of the existence of multiple-strain infections and
81 the propensity of HCMV to mutate during cell culture [6, 7, 8, 14, 15] added impetus to
82 sequencing genomes directly from clinical material in order to define natural populations.
83 One strategy for this involves sequencing overlapping PCR amplicons [7, 16]. Another
84 utilises an oligonucleotide bait library representing known HCMV diversity to select target
85 sequences from random DNA fragments. This target enrichment technology originated in
86 commercial kits for cellular exome sequencing, and was subsequently applied to various
87 pathogens [17, 18], including HCMV [19-21]. We have applied it since 2012 and have

88 systematically released via GenBank many genome sequences that have proved pivotal in
89 other studies [11, 12, 19-21].
90 The HCMV genome exhibits several evolutionary phenomena, including hypervariation,
91 multiple-strain infection, recombination and gene loss by mutation, all of which were
92 discovered prior to high-throughput sequencing and have since been illuminated by this
93 technology (early references are [22-26]). In this article, we explore these and other key
94 genomic features of HCMV, with an emphasis on the strains present in clinical material.

METHODS

Samples

For convenience, samples were analysed as three collections detailed in **Supplementary Tables 1-3** and summarised in **Table 1**, which also contains information on ethical permissions. 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 sequenced using target enrichment, but with a different oligonucleotide bait library. The features of the samples are in **Supplementary Tables 1-3** rows 3-6, and the clinical outcomes of congenital infection are in **Supplementary Table 1** row 205.

DNA sequencing

Target enrichment and sequencing library preparation were performed using the SureSelect^{XT} v. 1.7 system for Illumina paired-end libraries with biotinylated cRNA probe baits (Agilent, Stockport, UK) [21]. 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 generated using a MiSeq (Illumina, San Diego, CA, USA). Their names are in **Supplementary Tables 1-3** row 7. They were prepared for analysis using Trim Galore v 0.4.0

118 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/; length=21, quality=10
119 and stringency=3). The numbers of trimmed reads are in **Supplementary Tables 1-3** row 15.
120

121 **Library diversity**

122 The number of reads in a dataset that were derived from unique HCMV fragments was
123 estimated by using Bowtie2 v. 2.2.6 [29] to align the reads against the strain Merlin
124 sequence (GenBank accession AY446894.2), and also, where it could be determined, the
125 consensus genome sequence derived from the dataset. The relevant data are in
126 **Supplementary Tables 1-3** rows 17-19 and 23-26. Reads containing insertions or deletions,
127 duplicate read pairs sharing both end coordinates and duplicate unpaired reads sharing one
128 end coordinate were removed, thereby producing an alignment file for unique reads
129 derived from unique HCMV fragments (https://centre-for-virus-
130 research.github.io/VATK/AssemblyPostProcessing). This file was viewed using Tablet v.
131 1.14.11.7 [30]. The coverage depth values for total and unique fragment reads are in
132 **Supplementary Tables 1-3** rows 20-21 and 27-28.
133

134 **Strain enumeration**

135 The numbers of strains represented in each dataset were estimated by two strategies:
136 genotype read-matching and motif read-matching (https://centre-for-virus-
137 research.github.io/VATK/HCMV_pipeline). Both strategies utilised datasets concatenated
138 from the paired-end datasets. The genotype designations used were either based on
139 reported phylogenies [6, 12, 25, 31, 32], amended or extended as appropriate, or were
140 constructed afresh using Clustal Omega v. 1.2.4 [33] and MEGA v. 6.0.6 [34] with data for

141 the genomes listed in **Supplementary Table 4** and individual genes for which additional
142 sequences were available in GenBank. Alignments are in **Supplementary Figure 1**.
143 For genotype read-matching, Bowtie2 was used to align the reads to sequences
144 representing the genotypes of two hypervariable genes, UL146 and RL13 [6, 12, 35]. The
145 sequences representinng the entire coding region of UL146 and the central coding region of
146 RL13 are in **Supplementary Tables 1-3** rows 34-58. In contrast to the UL146 genotypes, the
147 RL13 genotypes cross-matched to some extent within four groups (G1, G2, G3; G4A, G4B;
148 G6, G10; and G7, G8). In these instances, the genotype within the group with most matching
149 reads was scored. The numbers of reads aligned to each genotype are in **Supplementary**
150 **Tables 1-3** rows 34-58. A genotype was scored if the number of reads was >10 and
151 represented >2% of the total number detected for all genotypes of that gene. For 14
152 samples in collection 1 that had been sequenced prior to the availability of ultrapure
153 (TruGrade) oligonucleotides, these values were set at >25 and >5%, respectively. The
154 number of strains in a sample was scored as the greater of the numbers of genotypes
155 detected for the two target genes, and is in **Supplementary Tables 1-3** row 13.
156 For motif read-matching, conserved genotype-specific motifs (20-31 nt) were identified
157 by visual inspection of the alignments for 12 hypervariable genes (**Supplementary Figure 1**)
158 [6, 12, 19, 33]. Additional motifs were included in order to identify the more common
159 intergenotypic recombinants and pseudogenic mutants. The motif sequences and number
160 of reads containing perfect matches to a sequence or its reverse complement are in
161 **Supplementary Tables 1-3** rows 60-178. Genotypes were scored as described above. The
162 number of strains in a sample was scored as the maximum number of genotypes detected
163 for at least two genes, and is in **Supplementary Tables 1-3** row 14.
164

165 **Data deposition**

166 The original read datasets were purged of human data and then deposited in the
167 European Nucleotide Archive (ENA; project no. PRJEB29585), and the consensus genome
168 sequences were deposited in GenBank. The accession numbers are in **Supplementary**
169 **Tables 1-3** rows 8 and 29, respectively. Updated genome sequence determinations in
170 collection 3 were deposited by the original submitters in GenBank [19] or by us as third-
171 party annotations in ENA (project no. PRJEB29374) [20]. Sequence features are in
172 **Supplementary Tables 1-3** rows 30-32. Data on pseudogenic mutants (obtained by motif
173 read-matching) in collection 1 are in **Supplementary Table 1** rows 180-203.

174

175 **Intrahost variant analysis**

176 Minor variant nucleotides were examined in datasets for which a consensus genome
177 sequence had been determined. Original datasets were prepared for analysis using Trim
178 Galore (length=100, quality=30 and stringency=1), and trimmed reads were mapped using
179 Bowtie2. Alignment files in SAM format were converted into BAM format, sorted using
180 SAMtools v. 1.3 [36], and analysed using LoFreq v. 2.1.2 [37] and V-Phaser 2 [38] under
181 default parameters.

182 **RESULTS**

183

184 **Operational limitations**

185 A total of 207 datasets from 199 samples and 102 individuals were analysed (**Table 1**
186 and **Supplementary Tables 1-3**). Library quality was represented in the percentage of HCMV
187 reads and the coverage depth by unique fragment reads. These values were related to
188 sample type, being higher for urine than blood presumably because of a higher proportion
189 of viral to host DNA. They also depended on the number of viral genome copies used to
190 make the library, with >1000 copies generally being needed to determine a complete
191 genome sequence. However, despite high library diversity, it was not possible to assemble
192 complete genome sequences from most datasets in collection 3 because of gaps in RL12 and
193 G+C-rich regions. This may have been because of limitations in the bait library and
194 sequencing approach used. The use of excessive PCR cycles with some samples in collections
195 1 and 2 led to high coverage depth by total fragment reads but low coverage depth by
196 unique fragment reads, and thus to highly clonal libraries (e.g. PAV2 in collection 1).
197 Genotypes present at subthreshold levels may represent multiple-strain infections or cross-
198 contamination during the complex sample processing pathway (e.g. PRA4 reads in PRA6
199 dataset 78943 in collection 1).

200

201 **Genome sequences**

202 A total of 91 complete or almost complete HCMV genome sequences were determined
203 (**Table 1**). We reported five previously [21], and 16 are improvements on published
204 sequences [19]. Most originated from single-strain infections or multiple-strain infections in
205 which one strain was predominant, and some originated from different strains that

206 predominated in a patient at different times. Defining a strain on the basis of a viral genome
207 present in an individual, these 91 sequences, plus an additional 49 deposited by our group
208 and 104 by other groups, brought the number of strains sequenced to 244 (**Supplementary**
209 **Table 4**). Of these, 91 were sequenced directly from clinical material, and all but one were
210 determined in this and our previous report [21]. The average size of the HCMV genome,
211 based on the 78 complete sequences in this set, is 235,465 bp (range 234,316-237,120 bp).
212

213 **Multiple-strain infections**

214 Genotypic differences in hypervariable genes (**Figure 1** and **Supplementary Figure 1**)
215 were exploited to detect multiple-strain infections by genotype read-matching and motif
216 read-matching , the latter method proving to be the more sensitive. Single strains were
217 common in congenitally infected patients (n=43/50 in collections 1 and 2), but significantly
218 less so in transplant recipients (n=11/25 in collections 2 and 3; chi-square test, χ^2=14.583,
219 p<0.05).
220

221 **Recombination**

222 The 244 genome sequences were genotyped in the 12 hypervariable genes used for
223 motif read-matching and then in five additional genes (**Figure 1** and **Supplementary Table**
224 **4**).

225 Hypervariation in UL55, which encodes glycoprotein B (gB), is located in two regions
226 (UL55N, near the N terminus, and UL55X, encompassing the proteolytic cleavage site) [23,
227 39]. Five genotypes (G1-G5) have been assigned to each region [23, 39-41], which are
228 separated by 927 bp that are 80% identical in all strains. All genomes had one of the
229 recognised UL55X genotypes (**Supplementary Table 5**). As reported previously [39], UL55N

230 G2 and G3 could not be distinguished reliably from each other, and two additional
231 genotypes (G6-G7) were detected that may have arisen from ancient recombination events
232 within UL55N (**Supplementary Tables 4** and **5**, and **Supplementary Figure 1**). There was
233 evidence for recombination in the region between UL55N and UL55X in only eight genomes.
234 This low proportion of recombination (3.3%) contrasts with the higher levels proposed
235 previously from PCR-based studies [39, 42].
236 UL73 and UL74, which encode glycoproteins N and O (gN and gO), respectively, are
237 adjacent hypervariable genes that exist as eight genotypes each [25, 32, 43]. There was
238 evidence for recombination between them in only seven genomes (2.9%), in accordance
239 with the low levels (2.2%) first detected in PCR-based studies [25, 32, 44]. In the region
240 containing adjacent hypervariable genes RL12, RL13 and UL1, recombinants were also rare
241 (1.2%) within RL12 and absent from RL13 and UL1. In contrast, hypervariable genes UL146
242 and UL139, which encode a CXC chemokine and a membrane glycoprotein, respectively, are
243 separated by a well-conserved region of over 5 kbp. The number (66) of the 126 possible
244 genotype combinations represented in the 244 genomes is too large to allow any underlying
245 genotypic linkage to be discerned, consistent with previous conclusions from PCR-based
246 studies [31]. No recombinants were noted within UL146.
247 In principle, strains in multiple-strain infections have the opportunity to recombine. In
248 our previous analysis of RTR1 in collection 2, we noted that one strain (RTR1A)
249 predominated at earlier times and another (RTR1B) at later times [21]. From the low
250 frequency of variants across a large part of the genome, we concluded that the second
251 strain had arisen either by recombination involving the first strain or by reinfection with, or
252 reactivation of, a second strain fortuitously similar to the first. In the present study,
253 recombination was strongly supported by a comparison of the two genome sequences,

254 which showed that approximately two-thirds of the genome is almost identical (differing by
255 three substitutions in noncoding regions), whereas the remaining third is highly dissimilar.
256 To investigate whether strains have been transmitted without recombination occurring,
257 identical genotypic constellations were identified among the 244 genomes (**Supplementary**
258 **Table 6**). This revealed the existence of 12 haplotype groups within which multiple strains
259 exhibit no signs of having recombined since diverging from their last common ancestor;
260 these are termed nonrecombinant strains below. As an incidental outcome, the two strains
261 in group 1 (PRA8 and CZ/3/2012), which were characterised in different studies, were
262 confirmed as having originated from the same patient, reducing the set of sequenced
263 strains to 243. The results from the other 11 groups suggest that nonrecombinant strains
264 have been circulating, some for periods sufficient to allow the accumulation of >100
265 substitutions. Application of an evolutionary rate estimated for herpesviruses (3.5×10^{-8}
266 substitutions/nt/year) [45] implies that these periods may have extended to many
267 thousands of years. The distribution of substitutions across the genome in highly divergent
268 groups 9 and 10 was examined in further detail. Group 9 (three strains) exhibited 135
269 differences, with the 50 that would affect protein coding distributed among 38 genes, and
270 group 10 (two strains) exhibited 138 differences, with the 38 that would affect protein
271 coding distributed among 27 genes. No obvious bias was observed towards greater diversity
272 in any particular gene or group of genes, including those in the hypervariable category. As
273 suggested by the lack of diversity within genotypes in comparison with the marked diversity
274 among them (**Supplementary Figure 1**), these findings fit with the view that intense
275 diversification of the hypervariable genes occurred early in human or pre-human history
276 [25, 31] and has long since ceased.
277

278 **Pseudogenes**

279 Mutations have been catalogued in some HCMV strains that cause premature
280 translational termination [7, 11, 12, 26]. They take the form of substitutions introducing in-
281 frame stop codons or ablating splice sites, or insertions, deletions or inversions causing
282 frameshifting or loss of entire protein-coding regions. However, evidence for these viral
283 pseudogenes was derived mostly from strains isolated in cell culture, and it was unclear to
284 what extent they are present in natural populations. For example, in a study reporting that
285 75% of strains are mutated [12], 157 mutations were identified in 101 strains, with all but
286 one passaged in cell culture and 35 confirmed by PCR of the clinical material. Nonetheless,
287 the distribution of mutations among the 91 strains sequenced in the present study from
288 clinical material is similar to that among strains isolated in cell culture (**Table 2** and
289 **Supplementary Table 4**), thus validating the earlier conclusions.

290 Among the strains sequenced from clinical material, 77% are mutated in at least one
291 gene (compared with 79% among all sequenced strains), and one is mutated in as many as
292 six genes (Pat_D in collection 3). The most frequently mutated genes are UL9, RL5A, UL1 and
293 RL6 (members of the RL11 family), US7 and US9 (members of the US6 gene family), and
294 UL111A (encoding viral interleukin-10) (**Supplementary Figure 1**). The likelihood that many
295 of these mutants are ancient is supported by the finding that all were detected at levels very
296 close to 100% in collection 1, and by previous observations identifying the same mutation in
297 different strains [7, 12]. In addition, there was evidence from the PAV6 datasets for
298 maternal transmission of a US7 mutant (**Supplementary Table 1**), and from PCR data (not
299 shown) for maternal transmission of a UL111A mutant to PAV16. Moreover, nine of the
300 groups of nonrecombinant strains contained mutants, and some of the mutations were
301 common to group members (**Supplementary Table 6**) and even to additional strains among

302 the 243, indicating that they had been transferred by recombination. These observations
303 again indicate the longevity of many of the pseudogenes and their propagation by
304 recombination. Focusing on the most common mutations, strains in which UL9, RL5A, UL1,
305 US9, US7 and UL111A were affected (singly or in combination) were, like strains that are not
306 mutated in any gene, transmitted in congenital infection and, in some cases, linked to
307 defects in neurological development (**Supplementary Table 1**).
308

309 **Intrahost diversity**

310 In addition to mutations indicative of pseudogenes, HCMV populations in individuals
311 may include minor variant nucleotides (single nucleotide polymorphisms). LoFreq and V-
312 Phaser analyses showed that single-strain infections contained markedly fewer variants
313 (median values of 60 and 140, respectively) than multiple-strain infections (median values of
314 2444 and 2955, respectively; **Figure 2**). The differences between the values for single- and
315 multiple-strain infections were significant (Kruskal-Wallis rank-sum test; LoFreq, χ^2=67.918,
316 p<2.2e^{-16}; V-Phaser, χ^2=63.536, p=1.6e^{-15}). Seven outliers in single-strain infections were
317 common to both analyses (in order of decreasing number of variants, RTR6B, CMV-37, RTR2,
318 CMV-35, CMV-38, ERR1279054 and PAV6), one was reported by LoFreq only (PAV21), and
319 four were reported by V-Phaser only (CMV-19, CMV-31, PRA6A and SCRT12).

320 **DISCUSSION**

321

322 Advances in high-throughput sequencing technology have made it possible to generate a
323 wealth of viral genome information directly from clinical material. However, operational
324 limitations should be registered in assessing the data. These include sample characteristics
325 (source, viral content and presence of multiple strains), confounding factors (logistical errors
326 and cross-contamination), design of the bait library and the sequencing protocol (ability to
327 enrich all strains and acquire data across the genome), and quality and extent of the
328 sequencing data (library diversity and coverage depth). Since perceived levels of intrahost
329 variation are particularly sensitive to these factors, we proceeded cautiously with this
330 aspect. However, as indicated in our earlier study [21], it is clear that the number of minor
331 variants in single-strain infections was markedly less than that in multiple-strain infections.
332 It was also far less than that reported by others in samples from congenital infections [16].
333 The factors listed above may have been responsible for the outliers observed in single-strain
334 infections (**Figure 2**). In our view, accurate estimates of the levels of intrahost variation in
335 single-strain infections are not available from the present and previous studies, and will
336 require sequencing and bioinformatic approaches that are demonstrably reliable, robust
337 and reproducible [46, 47].

338 Whole-genome analyses have confirmed the significant role of recombination during
339 HCMV evolution reported in numerous earlier studies [12, 19]. Recombination has occurred
340 over a very long period but nonetheless remains limited in extent, with surviving events
341 being more numerous in long regions, less numerous in short regions, and rare or absent in
342 hypervariable regions, as would be consistent with the role of homologous recombination.
343 Recombination frequency is potentially restricted in some circumstances by functional

344 interdependence within the same protein (e.g. gB) or possibly between separate proteins
345 (e.g. gN and gO [25, 30, 43]). However, it is not known whether differential homologous
346 recombination due to sequence relatedness is of general biological significance for the virus.
347 Also, strains have circulated that seem not to have recombined over many thousands of
348 years.

349 The extent to which recombinants arise and survive in individuals with multiple-strain
350 infections is a separate topic. Except where populations fluctuate significantly and are
351 sampled serially (e.g. RTR1 in collection 2), it is difficult to approach this using short-read
352 data, as they are based on PCR methodologies prone to generating recombinational
353 artefacts. Long- or single-read sequencing technologies and new bioinformatic tools are
354 needed in this area. Also, conclusions drawn from transplant recipients, who are
355 immunosuppressed and in whom HCMV populations may be diversified by transplantation
356 from HCMV-positive donors or selected with antiviral drugs, are unlikely to represent
357 naturally adapted cycles, such as those involving maternal transmission via breast milk [48].

358 The frequent identification of mutations giving rise to pseudogenes, and their
359 apparently long history, reveals an interesting aspect of HCMV microevolution. The
360 implication that some mutants have a selective advantage in certain circumstances may be
361 extended to their evident ability to be transmitted congenitally and to be associated with
362 neurological sequelae, probably in combination with specific host factors. The relevant
363 genes are involved, or are suspected to be involved, in immune modulation. They include
364 UL111A, which encodes viral interleukin-10 [49], UL40, which is involved in protecting
365 infected cells against NK cell lysis [50] via its cleaved signal peptide, in which mutations
366 occur, and UL9 , which bears a potential immunoglobulin-binding domain [2].

367 Modern approaches offer a powerful means for analysing HCMV genomes directly from
368 clinical material, with the important proviso that the data should be quality assessed and
369 interpreted in the context of the known evolutionary and biological characteristics of the
370 virus. Extensive high-throughput sequence data are likely to illuminate further the
371 epidemiology, pathogenesis and evolution of HCMV in clinical and natural settings, thus
372 facilitating the identification of virulence determinants and the development of new
373 interventions.
374
375 **NOTES**
376
377 ***Acknowledgements.*** The authors thank Florent Lasalle, Daniel Depledge and Judith Breuer
378 (University College London) for kindly providing unpublished collection 3 datasets and for
379 updating the associated genome sequences in GenBank, and Jenny Witthuhn (Hannover
380 Medical School) for excellent technical assistance.
381
382 ***Supplementary data.*** Supplementary materials are available at *The Journal of Infectious*
383 *Diseases* online. These consist of data provided by the authors to benefit the reader, are not
384 copyedited, and are the sole responsibility of the authors. Questions or comments should
385 be addressed to the corresponding author.
386
387 ***Financial support.*** This work was supported by the Medical Research Council (grant
388 numbers MC_UU_12014/3 and MC_UU_12014/12 to A. J. D.); the Wellcome Trust (grant
389 numbers 204870/Z/16/Z to A. J. D and WT090323MA to W. G. W. G.); the Ministry of Health

390 of the Czech Republic for conceptual development of research organization 00064203
391 (University Hospital, Motol, Prague, Czech Republic) to P. H.; the Fondazione Regionale per
392 la Ricerca Biomedica, Regione Lombardia (grant number FRRB 2015-043) to D. L.; the
393 Niedersächsische Ministerium für Wissenschaft und Kultur (grant COALITION – Communities
394 Allied in Infection) to T. G.; the Deutsche Forschungsgemeinschaft Collaborative Research
395 Centre 900 (core project Z1, grant number SFB-9001) to T. F. S.; and the German Center of
396 Infection Research TTU "Infections of the Immunocompromised Host" to T. G. and T. F. S.
397 Two authors (E. H. and A. D.) were supported by the Infection Biology graduate program of
398 Hannover Biomedical Research School.
399
400 ***Potential conflicts of interest:*** G. S. W. reports that his part in the submitted work was
401 completed prior to his present employment. G. W. G. W. reports a grant from the Wellcome
402 Trust. D. L. reports a grant from the Fondazione Regionale per la Ricerca Biomedica, Regione
403 Lombardia. T. G. reports grants from the German Federal Ministry of Education and
404 Research and from the Niedersächsische Ministerium für Wissenschaft und Kultur. P. H.
405 reports a grant from the Ministry of Health of the Czech Republic for the conceptual
406 development of University Hospital, Motol, Prague, Czech Republic, personal fees and non-
407 financial support from MSD and from Chimaerix, and personal fees from Dynex that are
408 outside the scope of the submitted work. T. F. S. reports grants from the Deutsche
409 Forschungsgemeinschaft Collaborative Research Centre 900 and from the German Federal
410 Ministry of Education and Research. A. J. D. reports grants from the Medical Research
411 Council and the Wellcome Trust. All other authors reported no conflicts of interest. All
412 authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

413 **REFERENCES**

414

415 1. Puchhammer-Stöckl E, Görzer I. Cytomegalovirus and Epstein-Barr virus subtypes – the
416 search for clinical significance. J Clin Virol **2006**; 36:239-48.
417 2. Chee MS, Bankier AT, Beck S, et al. Analysis of the protein-coding content of the
418 sequence of human cytomegalovirus strain AD169. Curr Top Microbiol Immunol **1990**;
419 154:125-69.
420 3. Dunn W, Chou C, Li H, et al. Functional profiling of a human cytomegalovirus genome.
421 Proc Natl Acad Sci U S A **2003**; 100:14223-8.
422 4. Murphy E, Yu D, Grimwood J, et al. Coding potential of laboratory and clinical strains of
423 human cytomegalovirus. Proc Natl Acad Sci U S A **2003**; 100:14976-81.
424 5. Sinzger C, Hahn G, Digel M, et al. Cloning and sequencing of a highly productive,
425 endotheliotropic virus strain derived from human cytomegalovirus TB40/E. J Gen Virol
426 **2008**; 89:359-68.
427 6. Dolan A, Cunningham C, Hector RD, et al. Genetic content of wild-type human
428 cytomegalovirus. J Gen Virol **2004**; 85:1301-12.
429 7. Cunningham C, Gatherer D, Hilfrich B, et al. Sequences of complete human
430 cytomegalovirus genomes from infected cell cultures and clinical specimens. J Gen Virol
431 **2010**; 91:605-15.
432 8. Dargan DJ, Douglas E, Cunningham C, et al. Sequential mutations associated with
433 adaptation of human cytomegalovirus to growth in cell culture. J Gen Virol **2010**;
434 91:1535-46.
435 9. Bradley AJ, Lurain NS, Ghazal P, et al. High-throughput sequence analysis of variants of
436 human cytomegalovirus strains Towne and AD169. J Gen Virol **2009**; 90:2375-80.

437 10. Jung GS, Kim YY, Kim JI, et al. Full genome sequencing and analysis of human
438 cytomegalovirus strain JHC isolated from a Korean patient. Virus Res **2011**; 156:113-20.
439 11. Sijmons S, Thys K, Corthout M, et al. A method enabling high-throughput sequencing of
440 human cytomegalovirus complete genomes from clinical isolates. PLoS One **2014**;
441 9:e95501.
442 12. Sijmons S, Thys K, Mbong Ngwese M, et al. High-throughput analysis of human
443 cytomegalovirus genome diversity highlights the widespread occurrence of gene-
444 disrupting mutations and pervasive recombination. J Virol **2015**; 89:7673-95.
445 13. Zhao F, Shen ZZ, Liu ZY, et al. Identification and BAC construction of Han, the first
446 characterized HCMV clinical strain in China. J Med Virol **2016**; 88:859-70.
447 14. Cha TA, Tom E, Kemble GW, Duke GM, Mocarski ES, Spaete RR. Human cytomegalovirus
448 clinical isolates carry at least 19 genes not found in laboratory strains. J Virol **1996**;
449 70:78-83.
450 15. Stanton RJ, Baluchova K, Dargan DJ, et al. Reconstruction of the complete human
451 cytomegalovirus genome in a BAC reveals RL13 to be a potent inhibitor of replication. J
452 Clin Invest **2010**; 120:3191-208.
453 16. Renzette N, Bhattacharjee B, Jensen JD, Gibson L, Kowalik TF. Extensive genome-wide
454 variability of human cytomegalovirus in congenitally infected infants. PLoS Pathog **2011**;
455 7:e1001344.
456 17. Melnikov A, Galinsky K, Rogov P, et al. Hybrid selection for sequencing pathogen
457 genomes from clinical samples. Genome Biol **2011**; 12:R73.
458 18. Depledge DP, Palser AL, Watson SJ, et al. Specific capture and whole-genome
459 sequencing of viruses from clinical samples. PLoS One **2011**; 6:e27805.

460 19. Lassalle F, Depledge DP, Reeves MB, et al. Islands of linkage in an ocean of pervasive
461 recombination reveals two-speed evolution of human cytomegalovirus genomes. Virus
462 Evol **2016**; 2:vew017.
463 20. Houldcroft CJ, Bryant JM, Depledge DP, et al. Detection of low frequency multi-drug
464 resistance and novel putative Maribavir resistance in immunocompromised pediatric
465 patients with cytomegalovirus. Front Microbiol **2016**; 7:1317.
466 21. Hage E, Wilkie GS, Linnenweber-Held S, et al. Characterization of human
467 cytomegalovirus genome diversity in immunocompromised hosts by whole-genome
468 sequencing directly from clinical specimens. J Infect Dis **2017**; 215:1673-83.
469 22. Chou SW, Dennison KM. Analysis of interstrain variation in cytomegalovirus glycoprotein
470 B sequences encoding neutralization-related epitopes. J Infect Dis **1991**; 163:1229-34.
471 23. Meyer-König U, Ebert K, Schrage B, Pollak S, Hufert FT. Simultaneous infection of healthy
472 people with multiple human cytomegalovirus strains. Lancet **1998**; 352:1280-1.
473 24. Rasmussen L, Geissler A, Winters M. Inter- and intragenic variations complicate the
474 molecular epidemiology of human cytomegalovirus. J Infect Dis **2003**; 187:809-19.
475 25. Mattick C, Dewin D, Polley S, et al. Linkage of human cytomegalovirus glycoprotein gO
476 variant groups identified from worldwide clinical isolates with gN genotypes,
477 implications for disease associations and evidence for N-terminal sites of positive
478 selection. Virology **2004**; 318:582-97.
479 26. Sekulin K, Görzer I, Heiss-Czedik D, Puchhammer-Stöckl E. Analysis of the variability of
480 CMV strains in the RL11D domain of the RL11 multigene family. Virus Genes **2007**;
481 35:577-83.
482 27. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its
483 applications to single-cell sequencing. J Comput Biol **2012**; 19:455-77.

484 28. Silva GG, Dutilh BE, Matthews TD, et al. Combining de novo and reference-guided
485 assembly with scaffold_builder. Source Code Biol Med **2013**; 8:23.
486 29. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods **2012**;
487 9:357-9.
488 30. Milne I, Stephen G, Bayer M, et al. Using Tablet for visual exploration of second-
489 generation sequencing data. Brief Bioinform **2013**; 14:193-202.
490 31. Bradley AJ, Kovács IJ, Gatherer D, et al. Genotypic analysis of two hypervariable human
491 cytomegalovirus genes. J Med Virol **2008**; 80:1615-23.
492 32. Bates M, Monze M, Bima H, Kapambwe M, Kasolo FC, Gompels UA; CIGNIS study group.
493 High human cytomegalovirus loads and diverse linked variable genotypes in both HIV-1
494 infected and exposed, but uninfected, children in Africa. Virology **2008**; 382:28-36.
495 33. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein
496 multiple sequence alignments using Clustal Omega. Mol Syst Biol **2011**; 7:539.
497 34. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary
498 Genetics Analysis version 6.0. Mol Biol Evol **2013**; 30:2725-9.
499 35. Davison AJ, Holton M, Dolan A, Dargan DJ, Gatherer D, Hayward GS. Comparative
500 genomics of primate cytomegaloviruses. In: Reddehase MJ, ed. Cytomegaloviruses: from
501 molecular pathogenesis to intervention. Vol. 1. Norwich, UK: Caister Academic Press,
502 **2013**.
503 36. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and
504 SAMtools. Bioinformatics **2009**; 25:2078-9.
505 37. Wilm A, Aw PP, Bertrand D, et al. LoFreq: a sequence-quality aware, ultra-sensitive
506 variant caller for uncovering cell-population heterogeneity from high-throughput
507 sequencing datasets. Nucleic Acids Res **2012**; 40:11189-201.

508 38. Yang X, Charlebois P, Macalalad A, Henn MR, Zody MC. V-Phaser 2: variant inference for
509 viral populations. BMC Genomics **2013**; 14:674.
510 39. Meyer-König U, Haberland M, von Laer D, Haller O, Hufert FT. Intragenic variability of
511 human cytomegalovirus glycoprotein B in clinical strains. J Infect Dis **1998**; 177: 1162-9.
512 40. Shepp DH, Match ME, Lipson SM, Pergolizzi RG. A fifth human cytomegalovirus
513 glycoprotein B genotype. Res Virol **1998**; 149:109-14.
514 41. Deckers M, Hofmann J, Kreuzer KA, et al. High genotypic diversity and a novel variant of
515 human cytomegalovirus revealed by combined UL33/UL55 genotyping with broad-range
516 PCR. Virol J **2009**; 6:210.
517 42. Haberland M, Meyer-König U, Hufert FT. Variation within the glycoprotein B gene of
518 human cytomegalovirus is due to homologous recombination. J Gen Virol **1999**;
519 80:1495-500.
520 43. Paterson DA, Dyer AP, Milne RS, Sevilla-Reyes E, Gompels UA. A role for human
521 cytomegalovirus glycoprotein O (gO) in cell fusion and a new hypervariable locus.
522 Virology **2002**; 293:281-94.
523 44. Yan H, Koyano S, Inami Y, et al. Genetic linkage among human cytomegalovirus
524 glycoprotein N (gN) and gO genes, with evidence for recombination from congenitally
525 and post-natally infected Japanese infants. J Gen Virol **2008**; 89:2275-9.
526 45. McGeoch DJ, Cook S, Dolan A, Jamieson FE, Telford EA. Molecular phylogeny and
527 evolutionary timescale for the family of mammalian herpesviruses. J Mol Biol **1995**;
528 247:443-58.
529 46. Xu C, Nezami Ranjbar MR, Wu Z, DiCarlo J, Wang Y. Detecting very low allele fraction
530 variants using targeted DNA sequencing and a novel molecular barcode-aware variant
531 caller. BMC Genomics **2017**; 18:5.

532 47. Illingworth CJR, Roy S, Beale MA, Tutill H, Williams R, Breuer J. On the effective depth of
533 viral sequence data. Virus Evol **2017**; 3:vex030.
534 48. Suárez NM, Musonda KG, Escriva E, et al. Multiple-strain infections of human
535 cytomegalovirus with high genome diversity are common in breast milk from HIV-
536 positive women in Zambia. BioRxiv **2018**; 493742. doi: https://doi.org/10.1101/493742.
537 49. McSharry BP, Avdic S, Slobedman B. Human cytomegalovirus encoded homologs of
538 cytokines, chemokines and their receptors: roles in immunomodulation. Viruses **2012**;
539 4:2448-70.
540 50. Prod'homme V, Tomasec P, Cunningham C, et al. Human cytomegalovirus UL40 signal
541 peptide regulates cell surface expression of the Natural Killer cell ligands HLA-E and
542 gpUL18. J Immunol **2012**; 188:2794-804.

543 **FIGURE LEGENDS**

544

545 **Figure 1**

546 Locations in the human cytomegalovirus strain Merlin genome of genes used for
547 genotyping. The genome consists of two unique regions, U_L (1325-194343 bp) and U_S
548 (197627-233108 bp), the former flanked by inverted repeats TR_L (1-1324 bp) and IR_L
549 (194344-195667 bp), and the latter flanked by inverted repeats IR_S (195090-197626 bp) and
550 TR_S (233109- 235646 bp). Protein-coding regions are indicated by colour-shaded arrows,
551 and noncoding RNAs as narrower, white-shaded arrows, with gene nomenclature below.
552 Introns are shown as narrow white bars. The 12 genes (RL5A, RL6, RL12, RL13, UL1, UL9,
553 UL11, UL73, UL74, UL120, UL146 and UL139) used for motif read-matching are coloured red
554 <dark grey *in greyscale version*>. Two of these genes (RL13 and UL146) were also used for
555 genotype read-matching. The additional five genes (UL20, UL33, UL37, UL55 and US9) used
556 to genotype genome sequences by alignment are coloured orange <mid grey *in greyscale*
557 *version*>. All other genes are coloured pink <light grey *in greyscale version*>.

558

559 **Figure 2**

560 Box-and-whisker graphs showing the number of variants detected in single-strain and
561 multiple-strain infections using (A) LoFreq and (B) V-Phaser. The graphs were created using
562 ggplot2 (https://ggplot2.tidyverse.org). Single-strain (n=134 and 131, respectively) and
563 multiple-strain datasets (n=29 and 29, respectively) for which consensus genome sequences
564 were derived were identified by motif read-matching. The total number of variants (2-50%
565 of the total) in each dataset was enumerated; length polymorphisms were not considered.
566 Each box (light grey for single strains and dark grey for multiple strains) encompasses the

first to third quartiles (Q1 to Q3) and shows the median as a thick line. For each box, the horizontal line at the end of the upper dashed whisker marks the upper extreme (defined as the smaller of Q3+1.5(Q3-Q1) and the highest single value), and the horizontal line at the end of the lower dashed whisker marks marks the lower extreme (the greater of Q1-1.5(Q3-Q1) and the lowest single value).

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 the European Nucleotide Archive (study PRJEB12814) [20].

[d]HCMV, human cytomegalovirus; viral 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

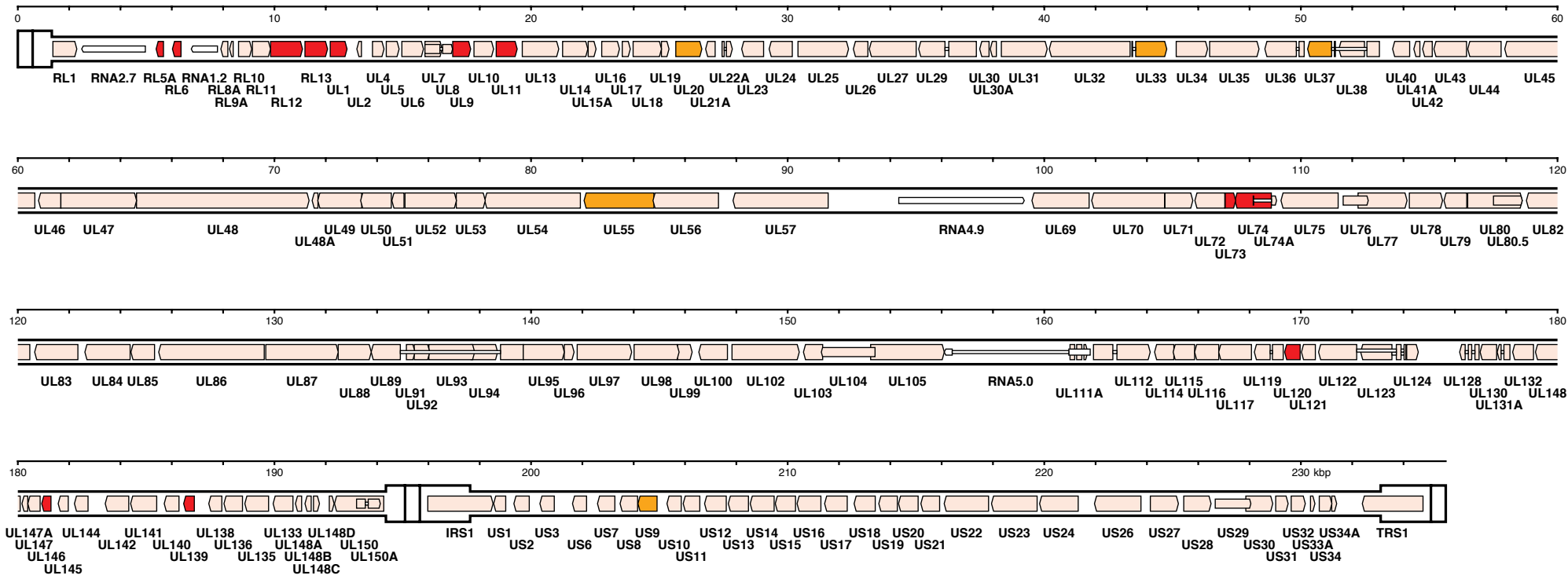
UL11	1	0	1	0.66	0.00	0.41
UL13	0	1	1	0.00	1.10	0.41
UL14	0	1	1	0.00	1.10	0.41
UL15A	0	1	1	0.00	1.10	0.41
UL20	1	0	1	0.66	0.00	0.41
UL43	0	1	1	0.00	1.10	0.41
UL99	1	0	1	0.66	0.00	0.41
UL148	1	0	1	0.66	0.00	0.41
UL147	1	0	1	0.66	0.00	0.41
UL145	0	1	1	0.00	1.10	0.41
UL150A	1	0	1	0.66	0.00	0.41
IRS1	1	0	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

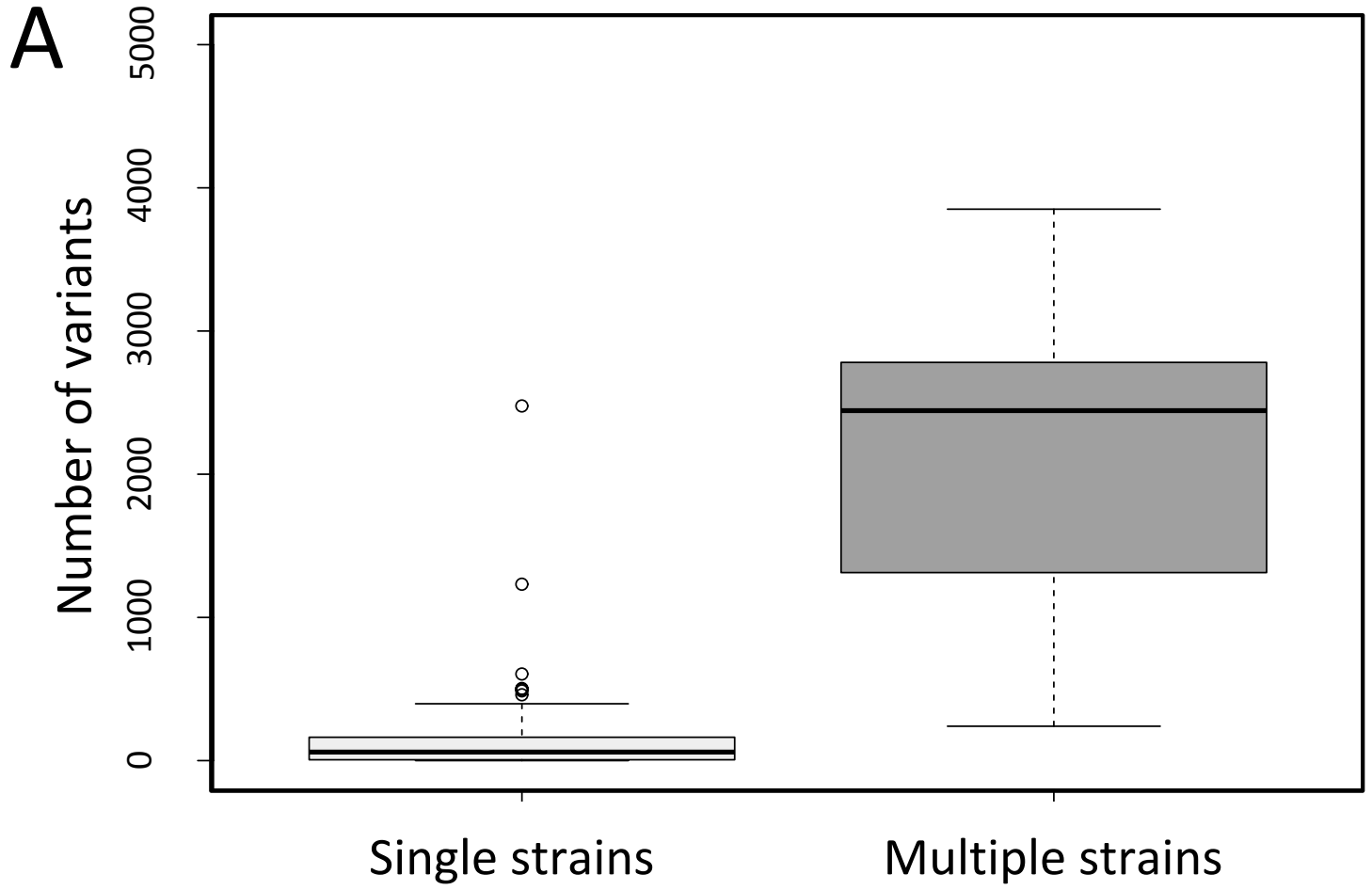
[a]Omitting mutations that occurred in RL13, UL128, UL130 and UL131A probably during passage, or that were engineered during bacterial artificial chromosome construction.

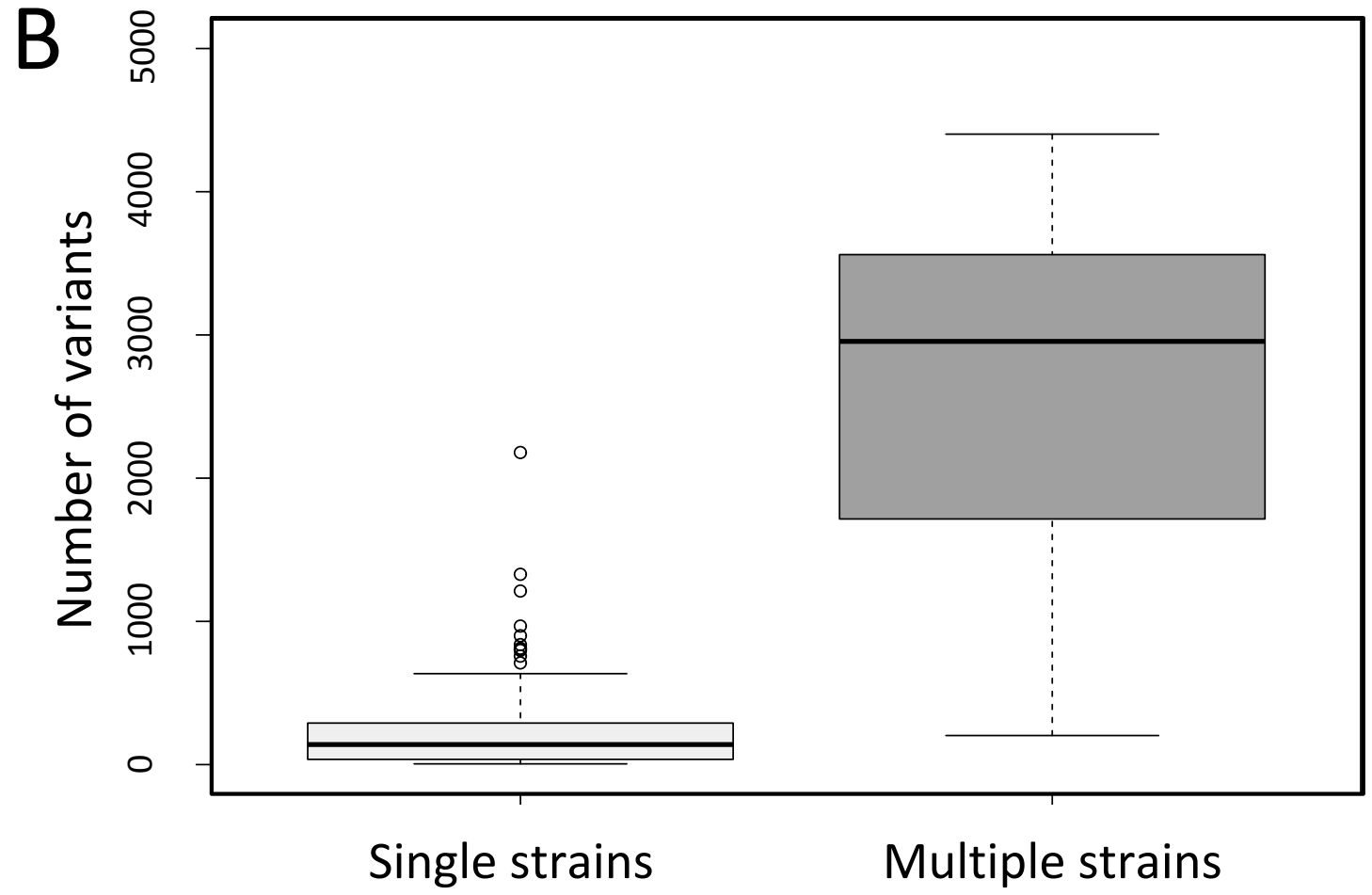
[b]Strains sequenced from strains passaged in cell culture, not taking into account the minority of mutations confirmed from the clinical samples (n=152, excludes CZ/3/2012, which is the same strain as PRA8).

[c]Strains sequenced directly from clinical material (n=91).

[d]Strains sequenced directly from clinical material or passaged virus (n=243).


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


A
Number of variants
0
1000
2000
3000
4000
5000
Single strains
Multiple strains


B
Number of variants
0
1000
2000
3000
4000
5000
Single strains
Multiple strains