

1 **Persistent low-level variants in a subset of HCMV genes are highly**
2 **predictive of poor outcome in immunocompromised patients with**
3 **cytomegalovirus infection**

4
5 Cristina Venturini^{1*}, Julia M Colston², Oscar Charles¹, Timothy Best³, Claire
6 Atkinson⁴, Calum Forrest⁴, Charlotte Williams⁵, Kanchan Rao⁶, Austen Worth⁷, Doug
7 Thorburn⁸, Mark Harber⁸, Paul Griffiths⁴ and Judith Breuer^{1,3}

8
9 ¹Infection, Immunity&Inflammation, Institute of Child Health, University College London, London,
10 UK

11 ²North Bristol NHS Trust, University Hospitals Bristol and Weston NHS Foundation Trust, Bristol,
12 UK

13 ³Great Ormond Street Hospital for Children NHS foundation trust, London, UK

14 ⁴Division of Infection and Immunity, Institute for Immunity and Transplantation, University College
15 London, London, UK

16 ⁵UCL Genomics, University College London, London, UK

17 ⁶Department of Bone Marrow Transplant, Great Ormond Street Hospital for Children NHS
18 foundation trust, London, UK

19 ⁷Department of Immunology, Great Ormond Street Hospital for Children NHS foundation trust,
20 London, UK

21 ⁸Royal Free London, NHS Foundation Trust, London, UK

22
23 *** Corresponding author:** Cristina Venturini (c.venturini@ucl.ac.uk)

24
25 **Abbreviations:**

26 HCMV – Human Cytomegalovirus

27 Kbp – kilobase pairs

28 NGS – Next Generation Sequencing

29 WGS – Whole Genome Sequencing

30 SOT – Solid organ transplant

31 HSCT - haematopoietic stem cell transplantation

32 PID – primary immunodeficiency

33 BMT – bone marrow transplant

34 GCV – ganciclovir
35 FOS – foscarnet
36 CDV – cidofovir
37 MV – minority variant

38

39 **ABSTRACT**

40

41 Human cytomegalovirus (HCMV) is the most common and most serious opportunistic infection
42 after solid organ (SOT) and haematopoietic stem cell transplantation (HSCT). There is considerable
43 interest in using virus sequence data to investigate and monitor viral factors associated with the
44 clinical outcome, including failure to respond to available antiviral therapies.

45 To assess this, we used target-enrichment to deep-sequence 16 paediatric patients with HSCT, SOT
46 or primary immunodeficiency of whom 9 died with HCMV and 35 infected SOT adult recipients of
47 whom one died with HCMV.

48 We first showed that samples from both groups have fixed drug-resistance mutations and mixed
49 infections. Deep sequencing also revealed non-fixed resistance mutations in most of the patients
50 who died (6/9). A machine learning approach identified 10 genes with high within-host variability in
51 these patients. These genes formed a viral signature which discriminated patients with HCMV who
52 died from those that survived with high accuracy (AUC=0.96). Lymphocyte numbers for a subset of
53 17 patients showed no recovery post-transplant of counts in the five who died.

54 We hypothesise that the viral signature identified in this study may be a useful biomarker for poor
55 response of HCMV to antiviral drug treatment and indirectly for poor T cell function, potentially
56 identifying early, those patients requiring non-pharmacological interventions.

57

58 **INTRODUCTION**

59

60 Human cytomegalovirus (HCMV; human herpesvirus 5) is a member of the *Betaherpesvirinae*
61 subfamily with a worldwide seroprevalence of between 18-100% (1,2). Higher prevalence is linked
62 to lower socio-economic status and older age. HCMV is usually a benign viral infection in
63 immunocompetent individuals, however, it has been shown to be a significant cause of morbidity
64 and mortality in immunosuppressed patients (3,4).

65 Given that HCMV is the most common and most serious opportunistic infection in these patients,
66 strategies for prevention as well as treatment are of paramount important for transplant clinical
67 success and outcome. Several therapies exist for prophylaxis, pre-emptive therapy and/or

68 treatment of HCMV (5). Treatment with ganciclovir (GCV), foscarnet (FOS), cidofovir (CDV) or
69 letermovir has improved outcomes (6–8), although late resistance often occurs (9). Despite
70 excellent outcomes for most haematopoietic stem cell transplants (SCT) and solid organ (SOT)
71 transplant recipients with HCMV, severe life threatening disease can develop in approximately 20%
72 (7) to 50% of cases (10). Increased use of next generation sequencing (NGS) has associated the
73 presence of fixed drug mutations and mixed infections with poorer outcomes (11–14). To further
74 investigate the pathogenesis of life-threatening HCMV in transplant recipients we analysed 51
75 patients with refractory HCMV viraemia defined as persisting with less than 0.5 log reduction for
76 three weeks or more despite antiviral treatment (15). Sixteen were HCMV-infected paediatric
77 patients with HCST, SOT or primary immunodeficiency, of whom 9 died with HCMV disease. The
78 other 35 were HCMV-infected SOT adult recipients, of whom one died with HCMV disease. Using a
79 machine learning approach, we identified an HCMV molecular signature which discriminates
80 between patients with different clinical outcome.

81

82 **RESULTS**

83

84 **Patients' characteristics**

85 We analysed two different cohorts: 16 retrospectively identified paediatric patients from Great
86 Ormond Street hospital for Children (GOSH) with primary immunodeficiency syndromes (PIDs) or
87 HSCTs (the characteristics of patients are shown in Table 1) and 35 adult SOT recipients from Royal
88 Free Hospital London. All had HCMV viraemia persisting with ≤ 0.5 log reduction despite antiviral
89 treatment for 21 days or longer, which has been defined as refractory (15).

90 No patients received prophylaxis against HCMV, although all SCTs received standard acyclovir
91 prophylaxis against alpha-herpesviruses. Pre-emptive antiviral treatment for HCMV was initiated at
92 first detection in the PIDs, when viraemia exceeded 1000 IU/ml in the SCT recipients and 3000 IU/ml
93 in the SOTs. First line therapy was ganciclovir in the SOTs and PIDs and foscarnet in the SCT
94 recipients.

95 We stratified patients into two groups: a poor outcome group, defined as those who died with
96 HCMV viraemia (n=9) and a good outcome group defined as patients who cleared their HCMV
97 (n=42).

98

99 **Deep sequencing metrics**

100 A total of 149 samples from 51 patients (1-9 samples per patient) were mapped to the HCMV
101 reference strain Merlin genome (NC.006273). Samples were included in further analysis if they
102 reached >95% coverage of the reference strain and a mean read depth (MRD) of >10x. MRDs ranged
103 from 10x to 1407x (mean 143x). Details of mapping statistics and quality are shown in
104 Supplementary Table 1.

105

106 **Drug resistance mutations**

107 Resistance to antiviral drugs has been associated with poor outcome. To examine this in our
108 patient cohort we annotated sequence data using a comprehensive database, derived from
109 published literature (16), to identify fixed resistance in the UL97 (serine/threonine protein kinase)
110 and the UL54 (DNA polymerase) genes, which are the targets of the anti-HCMV drugs GCV (UL97
111 and UL54) and FOS (UL54) used here. Fixed resistance (>50% frequency) mutations were identified
112 in 5/42 (11.9%) in the good outcome group and 3/9 (33.5%) in the poor outcome group ($X^2=2.57$, p-
113 value=.19).

114

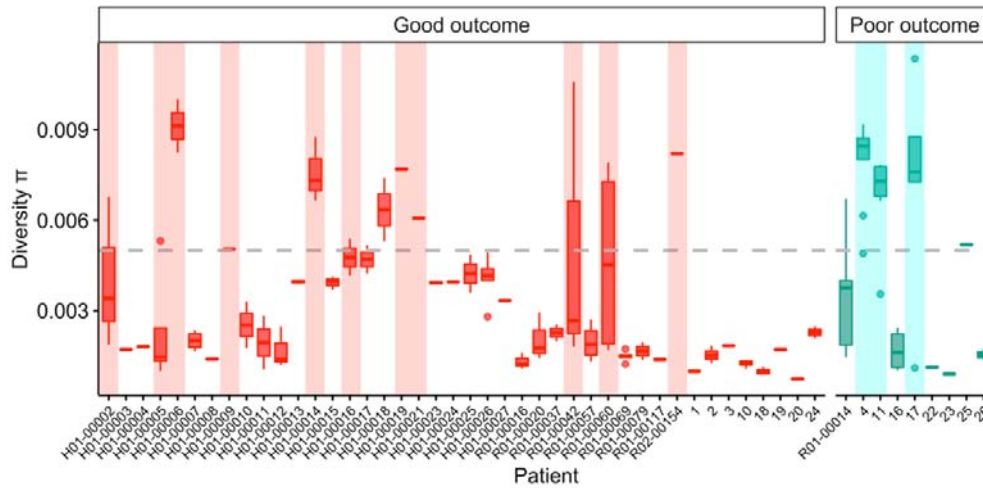
115 **Poor outcome is not associated with multiple HCMV strain infection**

116 Mixed HCMV infections have been previously linked with poor clinical outcome (12). To investigate
117 this in our two cohorts, we first calculated genome-wide within-host diversity (π) for each sample
118 (12) (Figure 1A). HCMV sample diversity separated into higher and lower diversity groups, with the
119 majority having low within-host diversity. The distribution of diversity showed bi-modality (p-
120 value=0.016, first peak/mode estimated at 0.0015 and second peak estimated at 0.0077, Figure 1 –
121 Supplementary figure 1A). The two estimated peaks were used to fit two Gaussian distributions to
122 the data (Figure 1 – Supplementary figure 1B) which crossed at >0.005. Based on previous analyses
123 (12) we used this value as the cut-off above which infections were considered as potentially mixed.
124 We reconstructed haplotypes in all 18 patients with at least one sample with diversity greater than
125 0.005. 14 patients were confirmed to have multiple strains where haplotypes differed by at least
126 2kbp with the minor haplotypes present at >5% frequency (17) (Figure 1B). Mixed infections were
127 not predictive of clinical outcome (26% of patients with good outcome presented with multiple
128 strains versus 33% with poor outcome, $X^2=0.19$, p-value=.66).

129

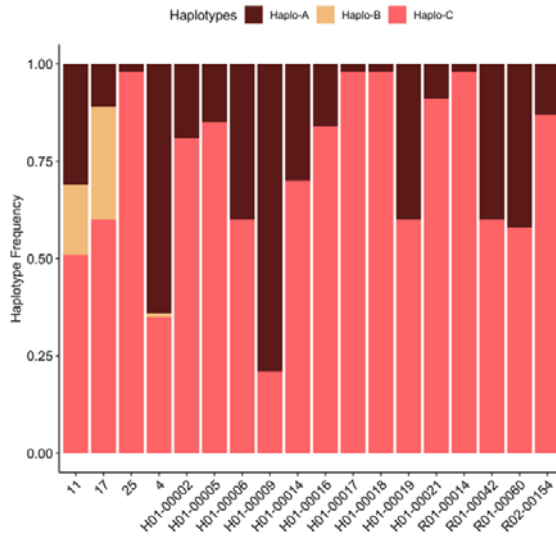
130

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).



131
132
133
134
135
136
137
138

Figure 1A: Diversity values for all patients. Each patient (x-axis) had longitudinal samples (box plots) showing diversity values (y-axis). Patients with good clinical outcomes are represented in red and those with poor outcomes are in turquoise. Dash grey line represents the chosen cut-off of 0.005 for investigating mixed infections. Patients with transparent rectangles were confirmed mixed infections.



139
140
141
142
143
144

Figure 1B: Reconstructed haplotype abundances for 18 patients suspected from their diversity scores of having mixed infections. For each patient haplotypes were reconstructed; here, for representative purposes, we only show one sample for each patient. Minor haplotypes occurring at <5% were discounted (17).

145

146 **Low-level variation in all HCMV genes and genes selection**

147 Deep sequencing also revealed low frequency (<50%) GCV and FOS resistance mutations in 6/9 of
 148 the poor outcome patients, two of whom also had mixed infections and 2/42 in the good prognosis
 149 group, none of whom had mixed infections (Table 2, $X^2=21.47$, p-value=.00001). Most variants
 150 occurred at frequencies <15% (median frequency 13.45) (Table 2 – figure supplement 1). In all
 151 patients in the poor outcome group with multiple longitudinal samples, low frequency resistance
 152 mutations persisted in multiple samples, with the majority failing to rise to fixation. In contrast, low
 153 level resistance variants present in the two patients in the good outcome group were at low
 154 frequency at the first-time point rising to fixation in later samples (Table 2 – figure supplement 1).

155

156

Clinical outcome	Patient	Fixed		Low level	
		UL54	UL97	UL54	UL97
Poor	P22	K513N, Q578L	M460I	E756D, Q578L ^a , A809V, L802M	
	P4			D588N, V715M	C592G, T409M, M460I
	P11			T813S, V715M	
	P25			E756D/Q, N408K, Q578H, L773V, A834P, G841A, A987G	C603W, H520Q, M460I, A594P
	P26		M460V		A594V
	R01-00014	L545S	M460I	L545S ^a	M460V
	3 patients (P16, P17 and P23) with poor clinical outcome did not show any resistance mutations				
Good	H01-00017	N408K	M460I	N408K ^a	

	H01-00016	L595S	
	H01-00012	C603W	C603W ^a
	H01-00003	L595S	
	P10	L501 [*]	G598D [*]
	37 patients with good clinical outcome did not show any resistance mutations		

157 **Table 2:** All detected drug resistance mutations (FC >=2)

158 ^{*}Detected only by Sanger sequencing

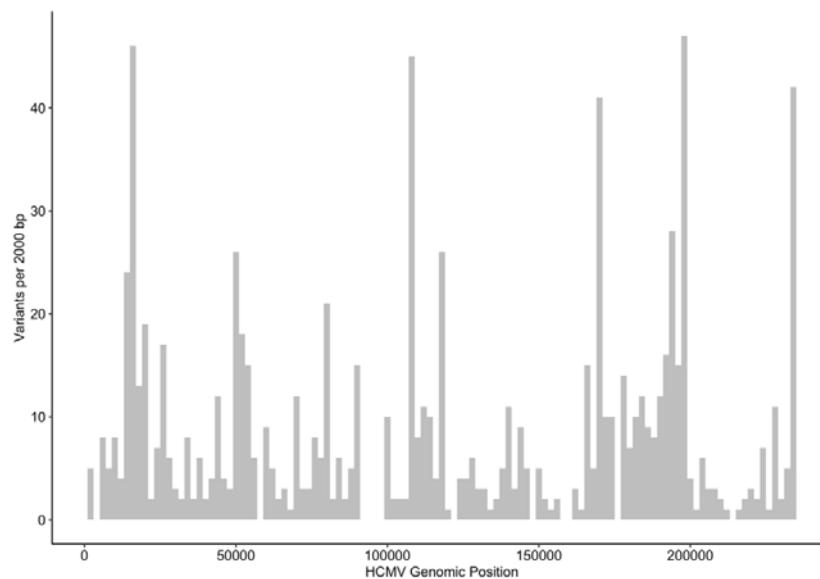
159 ^aVariants rising to fixation

160

161 To determine whether low level HCMV mutations (or minority variants, MVs) were confined to
162 positions coding for antiviral resistance, we investigated all non-synonymous (NS) MVs across the
163 genomes in the single infections (initially excluding all patients with mixed infection to simplify
164 mapping). In the whole population, there was no evidence of clustering (Figure 2) by open reading
165 frame, with NS MVs located apparently randomly across the HCMV genome.

166

167



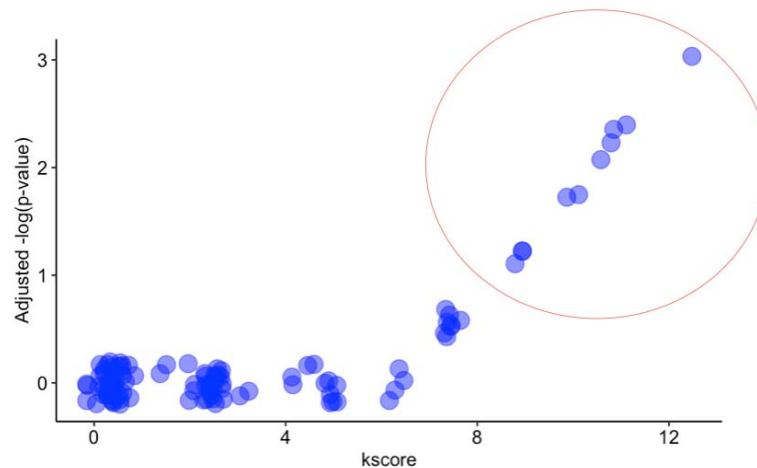
168

169 **Figure 2:** Minority variants distribution across the HCMV genome in all samples

170

171 We then compared samples from patients with single infections with different clinical outcomes.
172 We ranked the HCMV genes that discriminated patients who died from patients who survived using
173 machine learning methods. To do this, we implemented a gene selection algorithm (using chi-
174 squared statistics) to evaluate the importance of the presence of MVs in a specific gene. The gene

175 selection process identified 10 genes with high K score (K score > 8) and significant p-values (p-
176 values < 0.005 and adjusted p-value < 0.5) (Figure 3).
177
178



179
180 **Figure 3:** Feature selection results: x-axis shows the k-score and y-axis the -log(p-value) adjusted for multiple
181 testing. We selected the top 10 genes with k-score >8 and p-values < 0.005.

182
183 All ten genes showed more within-host variability in the poor outcome groups compared to the
184 good outcome group. No gene showed the opposite trend of being more variable in the good
185 outcome group compared to samples from patients who died.

186
187 The variable genes in the patients with poor outcome included the polymerase gene (UL54) and the
188 serine/threonine protein kinase (UL97), which are both already known for their association with
189 drug resistance. In addition, we identified genes coding for glycoproteins (envelope gp such as
190 UL74, gO and UL75, gH; immediate early gp, UL37; membrane gp UL7), membrane proteins UL121
191 and UL8 and the genes coding for the uncharacterized proteins UL20 and UL11, the latter of which
192 plays a role altering host immune response by modulating T-cell function.

193
194 We focussed on NS MVs as these gave better discrimination between poor and good outcome
195 groups than non-synonymous and synonymous mutations combined, for all genes, bar UL11, UL7
196 and UL97 (Table 3).

197

Ranking	Gene	NS K-score/p-value	All K-score/p-value
---------	------	--------------------	---------------------

1	UL54	12.5/4.07e-04	3/3.89e-03
2	UL20	11.2/8.18e-04	0.6/4.1e-01
3	UL11	10.6/1.08e-03	11/8.58e-04
4	UL8	10.6/1.08e-03	2/0.15
5	UL37	10.5/1.18e-03	0/1
6	UL121	10/1.56e-03	8/3.89e-03
7	UL75	10/1.56e-03	1.8/1.77e-01
8	UL7	8.8/2.98e-03	18/1.4e-05
9	UL97	8.8/2.98e-03	13/2.57e-04
10	UL74	8.7/3.09e-03	2/1.57e-01

198 **Table 3:** Genes are ranked by k-score and p-value. Column "All" shows k-score and p-values for analysis
199 including both synonymous and non-synonymous variants.

200

201 **Viral signature in HCMV samples from patients with poor clinical outcome**

202 To assess the predictive power of using the ten gene viral signature, we calculated the True Positive
203 Rate (TPR, sensitivity) and the False Positive Rate (FPR, 1-Specificity). We then built the Receiver
204 Operating Characteristics (ROC) curve and calculated the area under the curve (AUC) which
205 provides an aggregate measure of the performance of the model. We compared two models: a) a
206 full model including presence of NS MVs in the 10 signature's genes; and b) a drug resistance gene
207 model, where we only included NS MVs in UL54 and UL97. The full model had an AUC of 0.96 and
208 was statistically more discriminatory than the model with only resistance genes (p-value < 0.001,
209 Anova, Figure 4A). We used the full model to calculate, and plot estimated probabilities of having a
210 poor clinical outcome (Figure 4B).

211

212 Only 5 samples from 3 patients were misclassified by the full model (Table 4, Figure 4B). One
213 sample from the good outcome group was classified as "poor outcome". The patient (Ho1-00017)
214 was an adult who received a kidney transplant with two episodes of post-transplantation HCMV
215 viraemia (one lasting 28 days and one 149 days). Both samples were taken during the second
216 episode. The first sample taken at 77 days after transplant and 20 days from the start of the second
217 episode of HCMV viraemia and associated antiviral treatment, had a probability of 62% of poor
218 outcome. However, his second sample taken 25 days later, i.e. after 45 days of continuous
219 treatment, had a probability of poor outcome of 0%, despite the rise to fixation of a low level GCV
220 resistance mutation. This patient was in fact one of only two in the good outcome group with
221 multiple resistance mutations, one fixed and another rising to fixation in the second sample.

222

223 In the poor outcome group, patient P16 with RAG2 SCID, who died 45 days after starting antivirals
 224 for HCMV, had three of seven samples with a probability of <50% (28%, 40% and 27%) of being in
 225 the poor outcome group. These three samples were collected at days 3, 19 and 20 with viral loads of
 226 respectively 1016020, 174148 and 281838 gc/ml. Treatment with FOS and GCV was started at day
 227 17. The samples with low probability of poor outcome were interspersed with samples taken at days
 228 0 and 6 (viral loads 803282 and 553418 gc/ml respectively) showing probabilities of 90% of being in
 229 the poor outcome group. There was no correlation between viral load and probabilities, for example
 230 at day of peak viral load (1016020 gc/ml) the probability was 28%. A further two samples taken after
 231 day 20 showed probabilities of 76% and 100% of poor outcome. This patient did sadly subsequently
 232 die with HCMV infection.

233

234 A second patient (Ro1-00014) who died with HCMV infection also had an apparently low probability
 235 (2%) of poor outcome from a sample taken at 139 days post liver transplant. However, another
 236 three samples taken at 91, 153 and 171 days post-transplant showed probabilities of 100%, 72% and
 237 98% of poor outcome. Average read depth in all samples were >100x.

238

239 We also assessed the predictive power of the signature including single and mixed infections. The
 240 full model including mixed infections had a high predictive power (AUC=0.91), albeit lower than the
 241 model with only single infections (AUC=0.96), likely due to the difficulty in assembly and calling
 242 minority variants where two or more strains are present in a single sample (Figure 4 – supplement
 243 figure 1).

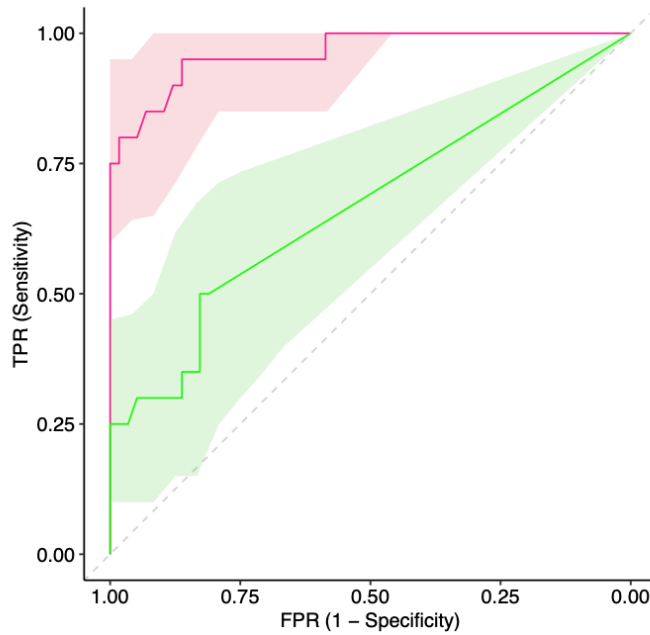
244

Patient	Longitudinal samples							
	Samples	Day 0	Day 3	Day 6	Day 19	Day 20	Day 52	Day 62
P16	Viral load (gc/ml)	803282	1016020	553418	174148	281838	282830	320073
	Probability	99%	28%	97%	40%	27%	76%	100%
	Samples	Day 91	Day 139	Day 153	Day 171			
Ro1-00014	Viral load (gc/ml)	11445	39218	512387	88500			
	Probability	100%	2%	72%	98%			
	Samples	Day 77		Day 122				
Ho1-00017	Samples	Day 77		Day 122				

	Viral load (gc/ml)	184783	46459
	Probability	62%	0%

245 **Table 4:** Patients with samples classified as “ambiguous”.

246



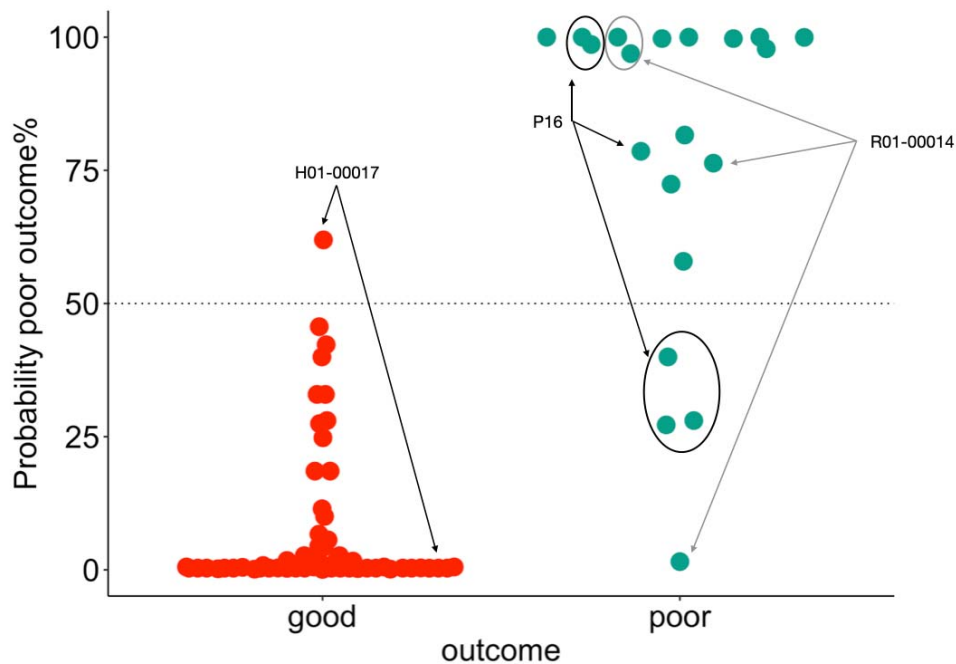
247

248 **Figure 4A:** ROC curves with confidence intervals (95%) for two predictive models discriminating between
249 samples from patients who died and survivors. AUC for the full model (including MVs in the 10 candidate
250 genes) was 0.96 (red ROC curve). AUC for the drug resistance genes model (including genes UL54 and UL97)
251 was 0.65 (green ROC curve).

252

253

254



255

256 **Figure 4B:** Estimated probabilities for each sample in the two groups (red for survivors, turquoise for patients
257 who died) to be classified as a patient from the poor outcome group. Arrows and circle indicate patients with
258 at least a sample which was misclassified by the model.

259

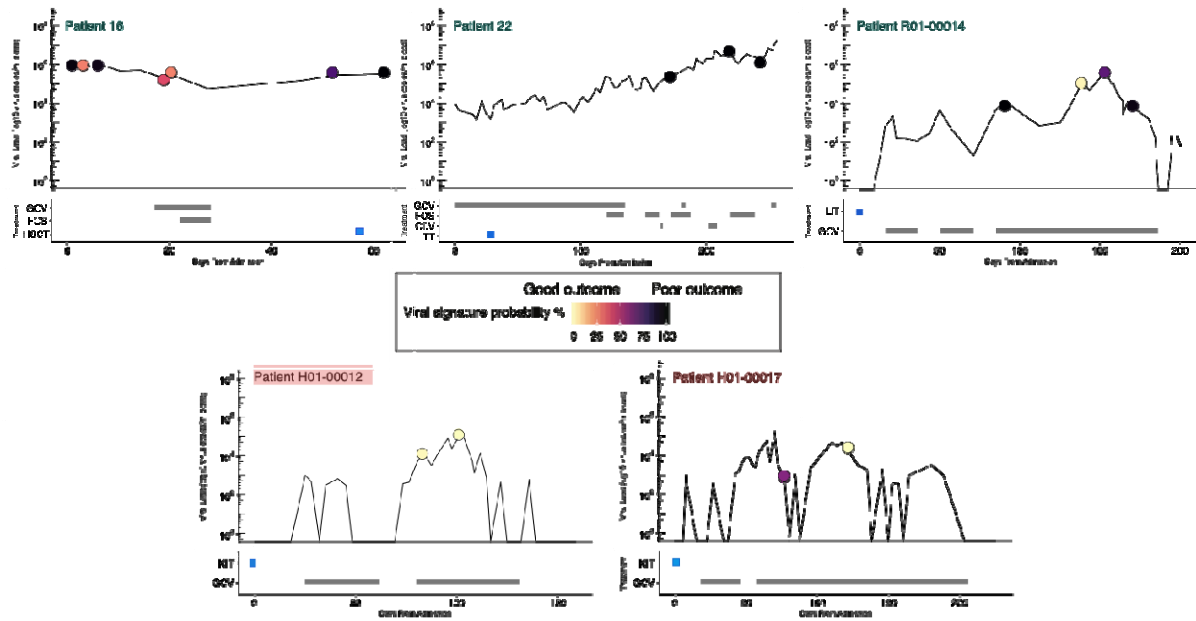
260 **Viral signature over time**

261 To determine how early low-level mutations in the ten sentinel genes can be used to predict a
262 potentially poor outcome, we plotted the probability of being in the poor outcome group for three
263 patients with poor outcome for whom we had multiple longitudinal samples (Figure 5). We also
264 plotted longitudinal data for two patients from the good outcome group who had low-level
265 resistance mutations.

266 Since samples from earlier during HCMV infection in patients 22 and R01-00014 with poor outcome
267 were not available for sequencing, it was not possible to determine the earliest time at which the
268 signature appeared. However, samples taken at days 171 (from admission) and 91 (from transplant)
269 respectively i.e. 62 and 109 days before death were positive for the predictive signature. In patient
270 P16 the signature was present as early as 9 days after HSCT.

271

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/) .



272

273 **Figure 5:** Viraemia, anti-viral therapy and transplant and viral signature probability in patients with
274 longitudinal samples and low-level resistance mutations, including patient 16, 22 and Ro1-00014 from the
275 poor outcome group and Ho1-00012 and Ho1-00017 from the good outcome group. Dots indicate samples
276 sequenced and are coloured based on viral signature probabilities (from yellow, good outcome, to black, poor
277 outcome). Black rectangles indicate anti-viral treatment, and the blue square shows the time of transplant
278 (HSCT for patient 16, liver for patient Ro1-00014 and thymus for patient 22, kidney for patients Ho1-00012
279 and Ho1-00017).

280

281 **Biological significance of the MVs**

282 Most of the HCMV genome is under purifying selection (18), presenting on average a greater
283 proportion of synonymous (S) changes compared to NS and stop codons. Surprisingly five of the
284 ten genes in our viral signature (UL54, UL20, UL121, UL97 and UL74) reversed this trend with
285 greater NS vs S MVs (Table 5). In the genes, NS variants tended to cluster closer together than
286 expected by chance suggesting a functional role. In addition, most of the MVs (63%) mapped to
287 HCMV variable loci identified comparing GenBank sequences. A higher overlap was observed for
288 hypervariable genes (e.g. UL74 (19)) compared with drug resistance genes (e.g. UL54) (Table 5).

289 Clustering of variable residues is a feature of epitopes for which plasticity provides advantages in
290 the face of host immunity. To explore this possibility, we identified known and predicted T cell
291 epitopes from the IEDB database which overlap with amino-acid changes seen in samples from
292 patients who died. We found epitopes in 8/10 genes, which included the 5 genes with greater NS vs
293 SMVs.

294

Gene	Description	% Of NS MVs in variable sites	NS - S	NS – S control	Do NS variants clustered together?	Epitopes
UL54	DNA polymerase catalytic subunit	0	33-13	10-22	p-value=3.4e-05	MLLDKEQM <u>A</u> LK; LE <u>N</u> GVTHRF; NHGAGG <u>T</u> AAVS YQGA
UL20	Uncharacterised	92%	57-42	13-21	p-value=3.6e-03	MLG <u>I</u> RAMLVMLDYYW; SS <u>T</u> E <u>G</u> NWSVTNLTES; MLL <u>P</u> RQYTL; FMDY <u>V</u> ILT <u>P</u> L <u>A</u> VLTC;
UL11	Plays a role in the modulation of host immune response by modulating T cell function	44%	34-39 2 stop codon	20-17	p-value=7.6e-04	CYYVVY <u>T</u> QNGTLPTT
UL8	Membrane protein	83%	47-47	23-35	p-value=6.5e-01	<u>S</u> SD <u>W</u> VTLGTS <u>A</u> S <u>L</u> LR
UL37	Immediate early glycoprotein	63%	52-52	26-27	p-value=2.8e-04	No epitope
UL121	Membrane protein	66%	11-7	10-10	p-value=1.4e-01	VCLILSFSIV <u>I</u> AALW; ISL <u>V</u> TPLTINATLRL; SCTHPYVISL <u>V</u> TPLT
UL75	Envelope glycoprotein gH	100%	25-63	8-18	p-value=9e-04	FPDATV <u>P</u> ATV; K <u>A</u> QLNRHSYLKDSDFLDA A; RQTEKHELLVLVKK <u>A</u> QLN RH; HELLVLVKK <u>A</u> QL; YLLSHLPSQRYGADAASE ALD <u>P</u> HAFHLLNTYGRPIR FLRENTTQC; A <u>A</u> SE <u>A</u> LD <u>P</u> HAFHLLNTY GR; LD <u>K</u> AFHLLL; YLL <u>S</u> HLP <u>S</u> QRYGAD <u>A</u> ASE ALDPHAFHLLNTYGRPIR FLRENTTQC
UL7	CEACAM1-like protein; plays a role in modulating the host immune	78%	20-33	13-20	p-value=3.5e-02	STPYVGLS <u>L</u> S <u>C</u> AANO

	response					
UL97	Serine/threonine protein kinase	11%	15 - 11	10 - 2	p-value=1.1e-02	No epitope
UL74	Envelope glycoprotein gO	92%	64 - 59	22-36	p-value=2.9e-05	LLFLDE <u>I</u> RNFSL <u>R</u> SP; TMRK <u>L</u> KRKQALVKEQ; SFYLVNAMSRLFRV

295 **Table 5:** The table shows biological features of the ten HCMV genes under investigation in patients who died.
 296 The table shows the % of NS variants mapping to HCMV variable sites, the number of NS vs S MVs found, p-
 297 values indicating whether the NS MVs clustered significantly closer than by chance and known and predicted
 298 T cell epitopes from the IEDB database which MVs mapped to (in bold and underline the position of the MV in
 299 the epitope).

300

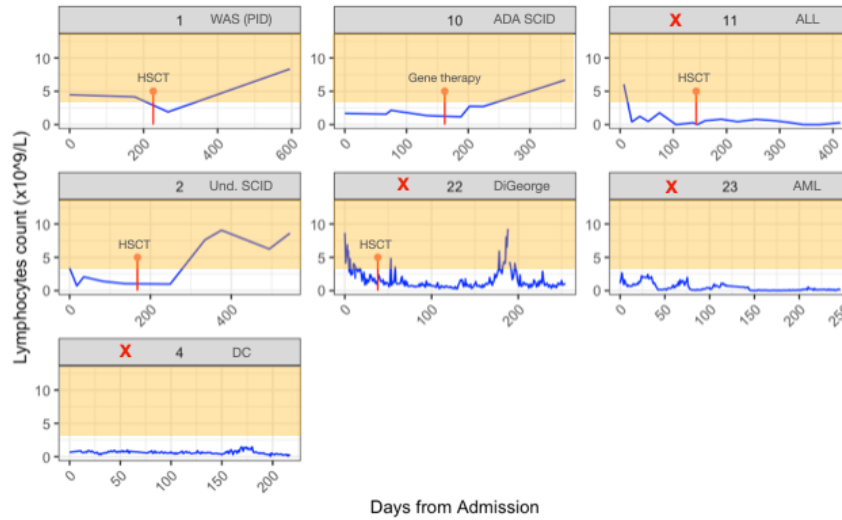
301 **Patients with poor clinical outcomes have lower lymphocyte counts**

302 The finding that MVs are significantly more likely to occur in regions predicted to be immunogenic
 303 led us to explore how immunity might relate to low level mutations. Lymphocyte counts were
 304 available from a subset of patients with PID/HSCT (Figure 6A). In patients 1 and 2 (who received
 305 SCTs) and patient 10 (who received gene therapy) in the good outcome group, lymphocyte counts
 306 recovered quickly after treatment (Figure 6A). In contrast, patient 4, 11, 22 and 23, who died,
 307 showed no recovery of lymphocyte count after HSCT. Lymphocyte counts were persistently low in
 308 both groups just after HSCT or gene therapy and started to increase at day 100 after transplant.
 309 Linear mixed effect modelling showed a significant difference in the counts over time (Figure 6B, p-
 310 value < .001) with significant differences in the final lymphocyte counts (good outcome median
 311 lymphocyte count: 8.34, 95%CI: 6.69- 8.34; poor outcome median lymphocyte count: 0.275, 95%CI:
 312 0.14-1.10).

313

314 Analysing the SOT adult cohort separately, as lymphocyte counts change with age, patient Ro1-
 315 00014 who died also showed persistently lower lymphocyte counts for months after receiving liver
 316 transplant as compared with the rest of the SOT cohort who survived (Figure 7A and 7B) (last time
 317 point before death for Ro1-00014 was 0.22, the median in the rest of the SOT patients was 1.61,
 318 95% 0.68-1.61).

319

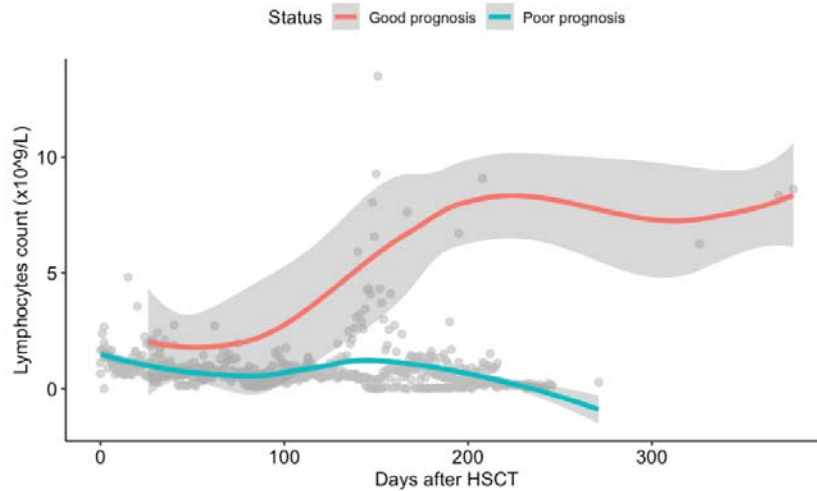


320

321 **Figure 6A:** Lymphocyte count (per microliter of blood) overtime in a subset of GOSH patients. Time of HSCT
322 or gene therapy is shown in red. In orange we indicate the healthy lymphocyte count range for children (3-13).

323 Patients with poor outcomes are indicated with a red cross.

324

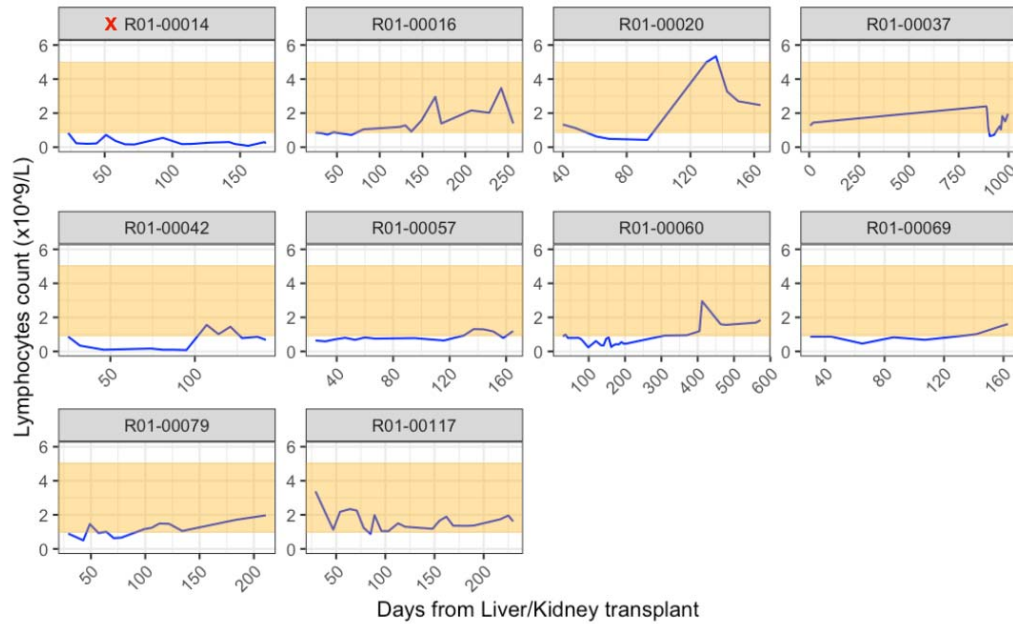


325

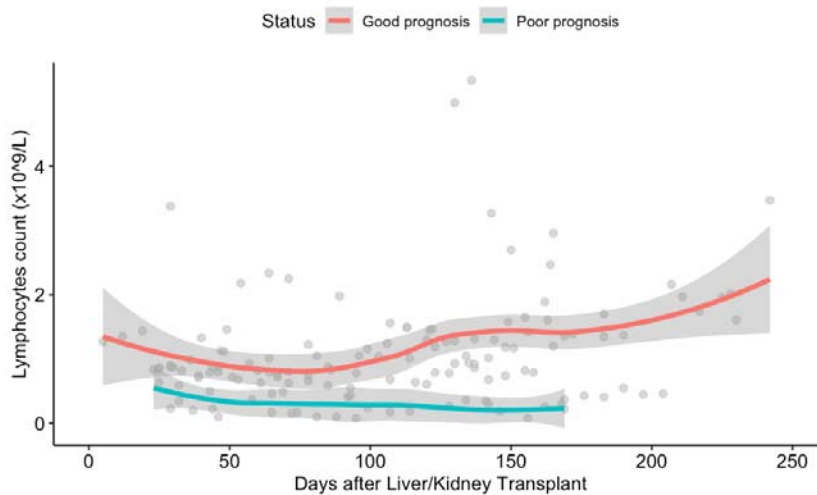
326 **Figure 6B:** Trend lines (smoothed local regression line using loess) for lymphocyte count for good and poor
327 outcome groups. The grey area represents 95% CI.

328

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).



329
330 **Figure 7A:** Lymphocyte count (per microliter of blood) overtime in a subset of liver/kidney adult patients. The
331 first time point is taken shortly after kidney/liver transplant. In orange we indicated the healthy lymphocyte
332 count range for adults (0.8-5).
333
334



335
336
337 **Figure 7B:** Trend lines (smoothed local regression line using loess) for lymphocyte count for good and poor
338 outcome groups in a subset of liver/kidney adult patients. The grey area represents 95% CI. X-axis is restricted
339 to 250 days after transplant.

340

341 **DISCUSSION**

342 Cytomegalovirus (HCMV) is the most common cause of infection following bone marrow and solid
343 organ transplants (7,20). Active disease significantly increases morbidity and mortality and
344 decreases graft survival in immunosuppressed patients. Pre-emptive therapy and prophylaxis have
345 reduced mortality, but HCMV related organ diseases remain a concern due to treatment side-
346 effects and the rise of drug resistant HCMV. Persistent viraemia is the outcome of the imbalance in
347 the interaction between the virus and the host: HCMV disease occurs when viral replication is
348 uncontrolled in a setting of impaired immune response (8). The mechanisms by which HCMV
349 infection influences transplant outcome are not known (7), but drug resistant HCMV strains and
350 infection with multiple strains have been associated with increased morbidity and mortality (11–14).

351 To investigate viral factors influencing transplant outcome and mortality, we deep sequenced a
352 total of 150 viruses from 16 children following SCT or SOT and from 35 adult SOT recipients. Nine
353 patients died with persistent HCMV viraemia, whilst the remaining 42 cleared their HCMV. Multiple-
354 strain infections are common in immune-compromised individuals and in our cohorts. We identified
355 a slightly higher percentage of mixed infections in patients who died (33% vs 26%), but the
356 difference was not significant in our study.

357 The use of antiviral drugs in the treatment of HCMV-disease perturbs the viral population, selecting
358 for variants in genes that cause drug-resistance. About 25% of the patients analysed in this study
359 showed resistance mutations at various levels in the DNA polymerase UL54 and the protein kinase
360 UL97, which are the major drug targets. Fixed mutations were present in both patients with poor
361 and good outcomes, albeit at lower levels in the latter. In contrast low level variants were almost
362 exclusively present in samples from patients who died. Interestingly, low level resistant variants
363 detected in two patients with good outcomes, quickly rose to fixation, whereas those detected in
364 three patients with poor outcomes for whom longitudinal samples were available persisted at low
365 level in the subsequent samples. Thus, the finding of low-level resistance mutations should trigger
366 repeat testing to better define the phenotype as well as to identify early resistance mutations that
367 may become fixed and require treatment change. Compared to traditional PCR and Sanger
368 sequencing, NGS can detect low-level resistance mutations at high resolution, enabling the
369 detection of evolving virus populations in immunocompromised individuals selected under anti-viral
370 treatment (11, 21–23).

371 These data and previous observations confirm that HCMV is highly stable at consensus level in
372 immunocompromised patients with very few substitutions observed over time in single-strain
373 infections (0-25 substitutions) (12,23,24). To further investigate the apparently greater low-level

374 mutability in some patients, we used a machine learning approach to attempt to discriminate
375 between patients with good and poor clinical outcomes. Comparing single-strain infected patients
376 who died with those who survived we identified the presence of low level (<50%) variants in one or
377 more of 10 genes, including UL54 and UL97 as discriminatory between the two groups.
378 Notwithstanding the opportunistic nature of the samples available, we were able to detect this
379 signature on average 84 days before death and <100 days post-transplant and in all cases the
380 signature was present in the first available sample from patients.

381 Most of the HCMV genome is under purifying selection (18), however 5/10 genes from the signature
382 identified had a higher proportion of NS variants than expected from pan- genome analysis and
383 mapped to loci known to be variable in HCMV. Variable loci clustered more than expected by
384 chance, hinting at the possibility of local positive selection a hallmark of immune epitopes (Table 5)
385 and indeed, in seven out of 10 of the signature protein genes MVs mapped to known HCMV T cell
386 epitopes. There might be several reasons why these variants remain at low frequency. Although
387 variation at consensus sequence level is rare due to the proofreading activity of the viral DNA
388 polymerase (25), one possibility is that low level variation in these epitopes occurs normally, but is
389 cleared by functional T cell immunity. Variants are unlikely to confer increased fitness, rising to
390 fixation only in circumstances where they enable evasion of prevailing immunity. In the absence of
391 functional T cell immunity, as in the cases with poor outcomes described here, we postulate that
392 variants arising in epitopes are able to persist at low level long enough to allow detection by deep
393 sequencing. Mutations arising in loci that confer resistance may rise to fixation in the presence of
394 the drug. However, there is evidence that GCV resistance mutations are not evenly distributed in
395 different cell compartments (26,27) and it is possible that the presence of low level virus
396 subpopulations with antiviral resistance represent virus sequestered in certain cell types.

397 Taken together the data hint at the possibility that dysregulated immunity in some way contributes
398 to the accumulation of low-level variants. There is evidence from early studies that recovery of
399 CD8+ T cells and CD4+ T cells is a positive predictor for prevention of mortality in HCMV-disease
400 (28–30). Restoration of HCMV-specific CTL response (class I MHC-restricted specific CD8+ CTL) may
401 require an extended time after transplant in some patients, and such patients are at increased risk
402 of developing severe HCMV disease. In our study, we were not able to obtain measurement of T cell
403 function, largely because the peripheral blood lymphocyte subset counts were too low for the
404 assays used. Instead, we analysed lymphocyte counts as proxy for lymphocyte function in a subset
405 of 7 children for whom data was available. None of the four patients (three post SCT and one with
406 PID) who died had measurable lymphocyte counts and all harboured viral variants in the ten genes

407 as described above. In contrast three subjects with a good outcome (two post HSCT and one after
408 gene therapy) for whom we had data showed good lymphocyte count recovery. Thus, the detection
409 of persistent low-level HCMV mutations following SCT may be a biomarker of poor immune
410 reconstitution and consequent poor outcome of HCMV infection. Although fatal HCMV disease is
411 less common in SOTs, it is interesting that patient Ro1-00014 who died with disseminated HCMV
412 showed a similar signature to the HSCT patients who died, suggesting that similar processes may
413 underlie fatal HCMV disease irrespective of transplant type. In this opportunistically collected
414 sample set, we did not always have samples early on in HCMV infection. Notwithstanding, the viral
415 signature was present, in all cases in the first sample tested, including at day nine following
416 transplant in one patient. In all cases the signature was detected <100 days after transplant, i.e.,
417 before T cell recovery is expected, thus providing a potential early biomarker for failure of
418 engraftment and poor outcome of HCMV infection. Since low level resistance mutations which later
419 rise to fixation can occur in the good prognosis group, repeated testing to demonstrate persistent
420 low level variants is likely to increase specificity. The treatment of patients with HCMV infections
421 that persist in the face of antiviral treatment is challenging. A biomarker that provides an early
422 indication of the likely failure of pharmacological approaches, could provide timely signposting of
423 the likely need for rescue cell-based therapies or even repeat HSCT to achieve control of HCMV in
424 these patients.

425 This study represents a set of preliminary observations based on a limited number of patients
426 especially in the poor outcome group. The findings now need to be tested prospectively in a larger
427 group of patients. A further limitation is that the biological basis for these observations is not
428 known although we speculate as to a possible explanation. Despite the availability of effective
429 antiviral prophylaxis and treatment, HCMV remains a serious infection, particularly in the context of
430 congenital or acquired persistent poor T cell numbers and function. In this context, the
431 development of antiviral resistance is more common, and the prognosis can be poor. Routine use
432 of next generation sequencing for HCMV resistance in refractory patients could potentially detect
433 significant resistance at earlier timepoints. At the same time, repeated detection of MVs may prove
434 to be a useful biomarker for poor response to drug treatment alone and identify patients, including
435 where there are insufficient cells present for functional T cell assays, for whom non-pharmaceutical
436 rescue therapies may be needed.

437

438 **METHODS**

439 **Sample collection and ethics**

440 *Great Ormond Street Hospital samples:*

441 Whole blood samples were stored at Great Ormond Street Hospital for Children (GOSH) at -80C.
442 These residual samples were collected as part of the standard clinical care at GOSH, and
443 subsequently approved for research use through the UCL Partners Infection DNA Bank by the NRES
444 Committee London Fulham (REC reference: 12/LO/1089) and West Midlands Black Country
445 Research Ethics Committee (REC reference: 18/WM/0186). All samples were anonymised. Informed
446 patient consent was not required. Nucleic acid was enriched using custom baits and sequenced as
447 previously described (11,12,31).

448

449 *Royal Free London samples:*

450 Samples were collected as part of the Wellcome collaborative grant 204870/Z/16/Z UKRI). UCL17-
451 0008 Analysis of Cytomegalovirus Pathogenesis in Solid Organ Transplant Patients approved by the
452 NRES Committee London Queens Square Ethics Committee (REC reference 17/LO/0916). Nucleic
453 acid was enriched using custom baits and sequenced as previously described (11,12,31).

454

455 **Data availability**

456 Data deposition: Raw sequencing data for HCMV have been deposited in the European Nucleotide
457 Archive (project accession no. PRJEB12814 and PRJEB55677 for GOSH patients and PRJEB55701 for
458 SOT WT patients).

459

460 **Statistical analysis**

461 *Bioinformatics analysis.* All sequences were analysed by the same methods. Reads were trimmed
462 and QC using Trimgalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and
463 then mapped to the Merlin strain (GenBank Id: NC_006273.2) using BBmap
464 (<https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/>). Variants were
465 called using Varscan version 2. All non-fixed variants were included in the analysis (frequency cut-
466 off 2%). Diversity calculations have been described elsewhere (12) and code is available here
467 <https://github.com/ucl-pathgenomics/NucleotideDiversity>. For haplotype reconstruction, we used
468 HaROLD, which uses co-varying variant frequencies in a probabilistic framework. Validation and
469 applications are described here (32) and programs can be found here [https://github.com/ucl-](https://github.com/ucl-pathgenomics/HaROLD)
470 [pathgenomics/HaROLD](https://github.com/ucl-pathgenomics/HaROLD).

471

472 *Feature selection.* We created a dataset where for each sample we had genes as categorical
473 variables and presence of MVs was indicated as 1/0. Genes with only one mutation in one sample

474 were filtered out. We implemented a gene selection algorithm to evaluate the importance of the
475 presence of low-level variants in a specific gene using Python scikit-learn library (33). Gene selection
476 was done according to the k highest scores (sklearn.feature_selection.SelectKBest with chi-square
477 statistics for classification). Data were split into train/test (70% train, 30% test) 1000 times and, to
478 avoid bias due to longitudinal sampling, we used a Leave-One-Out Cross-Validation (LOOCV)
479 procedure, in particular the shuffle-group-out cross-validation iterator implemented in scikit-learn
480 library (sklearn.model_selection.GroupShuffleSplit) which provides randomized train/test indices to
481 split data according to patient variable. Genes were selected based on 1) top 10 with the highest k-
482 score 2) adjusted $-\log(p\text{-value})$ of 1 and k-score of 8.

483

484 *Regression model accuracy and probability.* A generalised logistic model (R function glm, family
485 binomial) for implemented to test the accuracy of the 10-genes model in predicting the clinical
486 outcome. The 'predict()' function was used to predict the response value of each observations as
487 probability to be part of the poor prognosis group. ROC curves (function 'roc()') were used to show
488 the sensitivity/specificity for the binary classifier and the area under the curve (AUC) was also
489 calculated.

490 The results were compared with a model using only the two resistance genes, UL54 and UL97 using
491 an ANOVA test (likelihood-ratio, LR).

492

493 *Biology of the genes.* Genes were annotated and tested for drug resistance using the R package
494 cmvdr (16). T cell epitopes for HCMV were extracted from the Immune Epitope Database and
495 Analysis Resource (IEDB). Lymphocytes counts were compared with a mixed effect regression
496 model in R.

497

498 **FUNDING:**

499 CV, CA and CF are funded by the Wellcome Trust Grant No. "204870/Z/16/Z".

500 JB receives funding from the NIHR UCL/UCLH Biomedical Research Centre.

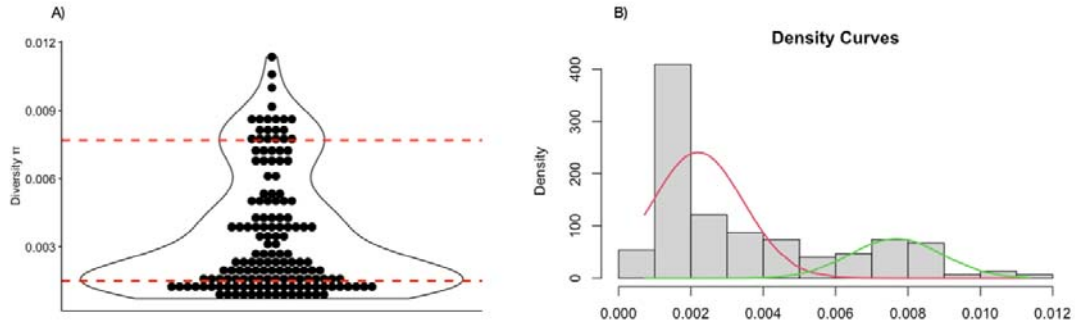
501

502 **ACKNOWLEDGMENTS:**

503 We are grateful for the excellent help from the Pathogen Genomics stream of UCL Genomics.

504

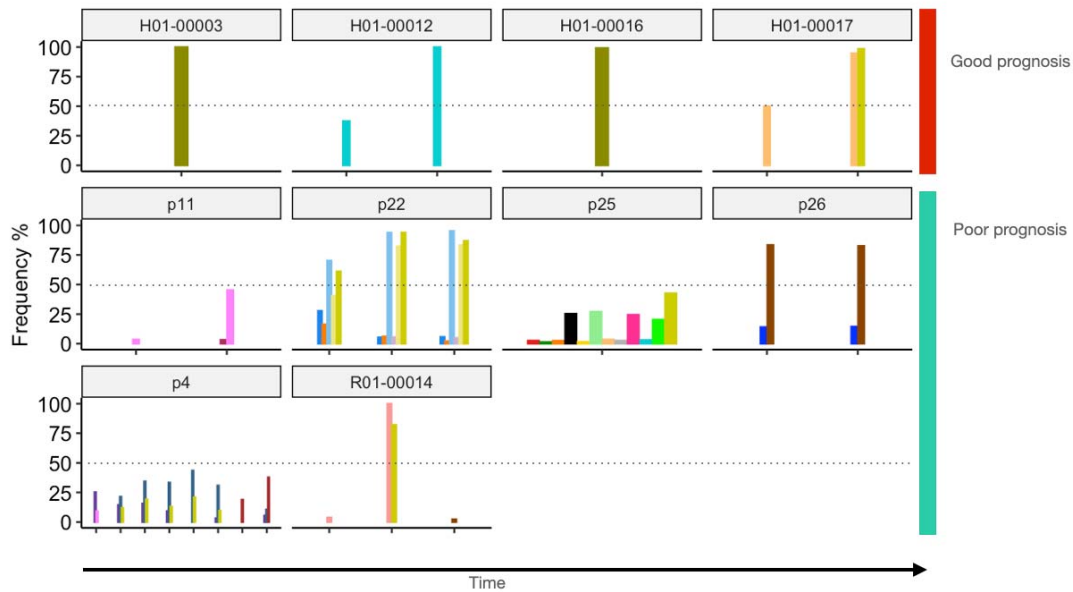
505 **SUPPLEMENTARY FIGURES:**



506

507 **Figure 1 – figure supplement 1:** Distribution of diversity values for all samples (A). Red dashed lines represent
508 the two modes of the bi-modal distribution. The estimated modes were used to create a mixture of Gaussian
509 distributions as shown in the plot (B).

510

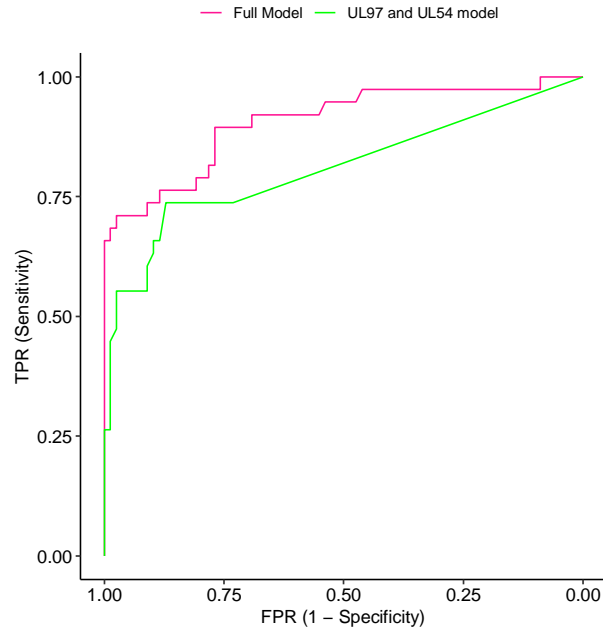


511

512

513 **Table 2 – figure supplement 1:** Resistance variants overtime (x-axis) in patients with good (red) and poor
514 (blue) clinical outcome. Variants are considered low-level if the frequency (y-axis) was below 50%.

515



516

517 **Figure 4 – supplement figure 1:** ROC curves with confidence intervals (95%) for two predictive models
518 discriminating between samples from patients who died and survivors including all samples from single and
519 mixed infections. AUC for the full model (including MVs in the 10 candidate genes) was 0.91 (red ROC curve).
520 AUC for the drug resistance genes model (including genes UL54 and UL97) was 0.81 (green ROC curve). The
521 two models were significantly different (p-value < 0.001, Anova).

522

523

524 **REFERENCES**

525

- 526 1. Cannon MJ, Schmid DS, Hyde TB. Review of cytomegalovirus seroprevalence and
527 demographic characteristics associated with infection. *Reviews in Medical Virology*.
528 2010;20(4):202–13.
- 529 2. Zuhair M, Smit GSA, Wallis G, Jabbar F, Smith C, Devleesschauwer B, et al. Estimation of the
530 worldwide seroprevalence of cytomegalovirus: A systematic review and meta-analysis.
531 *Reviews in Medical Virology*. 2019;29(3):e2034.
- 532 3. Green ML, Leisenring W, Xie H, Mast TC, Cui Y, Sandmaier BM, et al. Cytomegalovirus viral
533 load and mortality after haemopoietic stem cell transplantation in the era of pre-emptive
534 therapy: a retrospective cohort study. *The Lancet Haematology*. 2016 Mar 1;3(3):e119–27.
- 535 4. Ramanan P, Razonable RR. Cytomegalovirus Infections in Solid Organ Transplantation: A
536 Review. *Infect Chemother*. 2013 Sep 27;45(3):260–71.
- 537 5. Kotton CN, Kumar D, Caliendo AM, Huprikar S, Chou S, Danziger-Isakov L, et al. The Third
538 International Consensus Guidelines on the Management of Cytomegalovirus in Solid-organ
539 Transplantation. *Transplantation*. 2018 Jun;102(6):900–31.
- 540 6. Boeckh M, Ljungman P. How we treat cytomegalovirus in hematopoietic cell transplant
541 recipients. *Blood*. 2009 Jun 4;113(23):5711–9.
- 542 7. Kotton CN. Management of cytomegalovirus infection in solid organ transplantation. *Nat Rev*
543 *Nephrol*. 2010 Dec;6(12):711–21.
- 544 8. Bruminhent J, Razonable RR. Advances in drug therapies for cytomegalovirus in
545 transplantation: a focus on maribavir and letermovir. *Expert Opinion on Orphan Drugs*. 2020
546 Oct 2;8(10):393–401.
- 547 9. Shmueli E, Or R, Shapira MY, Resnick IB, Caplan O, Bdolah-Abram T, et al. High Rate of
548 Cytomegalovirus Drug Resistance Among Patients Receiving Preemptive Antiviral Treatment
549 After Haploidentical Stem Cell Transplantation. *The Journal of Infectious Diseases*. 2014 Feb
550 15;209(4):557–61.
- 551 10. van der Beek MT, Marijt EW, Vossen AC, van der Blij-de Brouwer CS, Wolterbeek R, Halkes CJ,
552 et al. Failure of Pre-Emptive Treatment of Cytomegalovirus Infections and Antiviral
553 Resistance in Stem Cell Transplant Recipients. *Antiviral Therapy*. 2012 Jan 1;17(1):45–51.
- 554 11. Houldcroft CJ, Bryant JM, Depledge DP, Margetts BK, Simmonds J, Nicolaou S, et al.
555 Detection of Low Frequency Multi-Drug Resistance and Novel Putative Maribavir Resistance
556 in Immunocompromised Pediatric Patients with Cytomegalovirus. *Front Microbiol* [Internet].
557 2016 [cited 2020 May 27];7. Available from:
558 <https://www.frontiersin.org/articles/10.3389/fmicb.2016.01317/full>
- 559 12. Cudini J, Roy S, Houldcroft CJ, Bryant JM, Depledge DP, Tutill H, et al. Human
560 cytomegalovirus haplotype reconstruction reveals high diversity due to superinfection and
561 evidence of within-host recombination. *PNAS*. 2019 Mar 19;116(12):5693–8.

- 562 13. Coaquette A, Bourgeois A, Dirand C, Varin A, Chen W, Herbein G. Mixed Cytomegalovirus
563 Glycoprotein B Genotypes in Immunocompromised Patients. *Clinical Infectious Diseases*.
564 2004 Jul 15;39(2):155–61.
- 565 14. Lisboa LF, Tong Y, Kumar D, Pang XL, Åsberg A, Hartmann A, et al. Analysis and clinical
566 correlation of genetic variation in cytomegalovirus. *Transplant Infectious Disease*.
567 2012;14(2):132–40.
- 568 15. Chemaly RF, Chou S, Einsele H, Griffiths P, Avery R, Razonable RR, et al. Definitions of
569 Resistant and Refractory Cytomegalovirus Infection and Disease in Transplant Recipients for
570 Use in Clinical Trials. *Clinical Infectious Diseases*. 2019 Apr 8;68(8):1420–6.
- 571 16. Charles OJ, Venturini C, Breuer J. cmvdrg - An R package for Human Cytomegalovirus antiviral
572 Drug Resistance Genotyping [Internet]. bioRxiv; 2020 [cited 2022 Apr 5]. p.
573 2020.05.15.097907. Available from:
574 <https://www.biorxiv.org/content/10.1101/2020.05.15.097907v1>
- 575 17. Pang J, Slyker JA, Roy S, Bryant J, Atkinson C, Cudini J, et al. Mixed cytomegalovirus
576 genotypes in HIV-positive mothers show compartmentalization and distinct patterns of
577 transmission to infants. Stanley M, Akhmanova A, Ramchandrar N, editors. *eLife*. 2020 Dec
578 31;9:e63199.
- 579 18. Lassalle F, Depledge DP, Reeves MB, Brown AC, Christiansen MT, Tutill HJ, et al. Islands of
580 linkage in an ocean of pervasive recombination reveals two-speed evolution of human
581 cytomegalovirus genomes. *Virus Evolution*. 2016 Jan 1;2(1):vew017.
- 582 19. Suárez NM, Musonda KG, Escriva E, Njenga M, Agbueze A, Camiolo S, et al. Multiple-Strain
583 Infections of Human Cytomegalovirus With High Genomic Diversity Are Common in Breast
584 Milk From Human Immunodeficiency Virus–Infected Women in Zambia. *The Journal of*
585 *Infectious Diseases*. 2019 Jul 31;220(5):792–801.
- 586 20. Hiwarkar P, Gaspar HB, Gilmour K, Jagani M, Chiesa R, Bennett-Rees N, et al. Impact of viral
587 reactivations in the era of pre-emptive antiviral drug therapy following allogeneic
588 haematopoietic SCT in paediatric recipients. *Bone Marrow Transplant*. 2013 Jun;48(6):803–8.
- 589 21. Chou S, Ercolani RJ, Sahoo MK, Lefterova MI, Strasfeld LM, Pinsky BA. Improved Detection of
590 Emerging Drug-Resistant Mutant Cytomegalovirus Subpopulations by Deep Sequencing.
591 *Antimicrob Agents Chemother*. 2014 Aug;58(8):4697–702.
- 592 22. Guermouche H, Burrell S, Mercier-Darty M, Kofman T, Rogier O, Pawlotsky JM, et al.
593 Characterization of the dynamics of human cytomegalovirus resistance to antiviral drugs by
594 ultra-deep sequencing. *Antiviral Research*. 2020 Jan 1;173:104647.
- 595 23. Suárez NM, Blyth E, Li K, Ganzenmueller T, Camiolo S, Avdic S, et al. Whole-Genome
596 Approach to Assessing Human Cytomegalovirus Dynamics in Transplant Patients Undergoing
597 Antiviral Therapy. *Front Cell Infect Microbiol* [Internet]. 2020 [cited 2020 Aug 18];10. Available
598 from: <https://www.frontiersin.org/articles/10.3389/fcimb.2020.00267/full#h3>
- 599 24. Hage E, Wilkie GS, Linnenweber-Held S, Dhingra A, Suárez NM, Schmidt JJ, et al.
600 Characterization of Human Cytomegalovirus Genome Diversity in Immunocompromised
601 Hosts by Whole-Genome Sequencing Directly From Clinical Specimens. *The Journal of*
602 *Infectious Diseases*. 2017 Jun 1;215(11):1673–83.

- 603 25. Renzette N, Gibson L, Jensen JD, Kowalik TF. Human cytomegalovirus intrahost evolution—a
604 new avenue for understanding and controlling herpesvirus infections. *Current Opinion in*
605 *Virology*. 2014 Oct 1;8:109–15.
- 606 26. Eckle T, Prix L, Jahn G, Klingebiel T, Handgretinger R, Selle B, et al. Drug-resistant human
607 cytomegalovirus infection in children after allogeneic stem cell transplantation may have
608 different clinical outcomes. *Blood*. 2000 Nov 1;96(9):3286–9.
- 609 27. Frange P, Boutolleau D, Leruez-Ville M, Touzot F, Cros G, Heritier S, et al. Temporal and
610 spatial compartmentalization of drug-resistant cytomegalovirus (CMV) in a child with CMV
611 meningoencephalitis: Implications for sampling in molecular diagnosis. *Journal of Clinical*
612 *Microbiology*. 2013;51(12):4266–9.
- 613 28. Griffiths P, Reeves M. Pathogenesis of human cytomegalovirus in the immunocompromised
614 host. *Nat Rev Microbiol*. 2021 Dec;19(12):759–73.
- 615 29. Reusser P, Riddell S, Meyers J, Greenberg P. Cytotoxic T-lymphocyte response to
616 cytomegalovirus after human allogeneic bone marrow transplantation: pattern of recovery
617 and correlation with cytomegalovirus infection and disease. *Blood*. 1991 Sep 1;78(5):1373–80.
- 618 30. Quinnan GV, Kirmani N, Rook AH, Manischewitz JF, Jackson L, Moreschi G, et al. Cytotoxic T
619 Cells in Cytomegalovirus Infection. *New England Journal of Medicine*. 1982 Jul 1;307(1):7–13.
- 620 31. Depledge DP, Palser AL, Watson SJ, Lai IYC, Gray ER, Grant P, et al. Specific Capture and
621 Whole-Genome Sequencing of Viruses from Clinical Samples. *PLOS ONE*. 2011 Nov
622 18;6(11):e27805.
- 623 32. Pang J, Venturini C, Tamuri AU, Roy S, Breuer J, Goldstein RA. Haplotype assignment of
624 longitudinal viral deep-sequencing data using co-variation of variant frequencies. *bioRxiv*.
625 2020 Aug 27;444877.
- 626 33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn:
627 Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12(85):2825–30.
- 628
- 629
- 630
- 631