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

Cristina Venturini1,\*, Julia M Colston2, Oscar Charles1, Anastasia Lankina1, Timothy Best3, Claire Atkinson4,5, Calum Forrest5, Charlotte A. Williams6, Kanchan Rao7, Austen Worth8, Doug Thorburn9, Mark Harber9, Paul Griffiths5 and Judith Breuer1,3

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

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

3 Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK

4 Applied Science, London South Bank University, London, UK

5 Division of Infection and Immunity, Institute for Immunity and Transplantation, University College London, London, UK

6 Department of Genetics and Genomic Medicine, GOS Institute of Child Health, University College London, London, UK.

7 Department of Bone Marrow Transplant, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK

8 Department of Immunology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK

9 Royal Free London, NHS Foundation Trust, London, UK

**\* Corresponding author:** Cristina Venturini PhD, Infection, Immunity & Inflammation, Institute of Child Health, University College London, WC1N 1EH, London, United Kingdom (c.venturini@ucl.ac.uk)

**Keywords:** cytomegalovirus infection, genomics, viral signature, transplant

**Short title:** HCMV genetic markers of outcome

**Short summary:**

In this study, we explore the complex relationship between cytomegalovirus infections and clinical outcomes in transplant patients and propose strategies for monitoring viral dynamics to optimise treatment and prevent failure to respond to anti-viral therapies.

**ABSTRACT**

**Background**

Human cytomegalovirus is the most common and serious opportunistic infection after solid organ and haematopoietic stem cell transplantation. In this study, we used whole-genome cytomegalovirus data to investigate viral factors associated with the clinical outcome.

**Methods**

We sequenced cytomegalovirus samples from 16 immunocompromised paediatric patients with persistent viraemia. 8/16 patients died of complications due to cytomegalovirus infection. We also sequenced samples from 35 infected solid organ adult recipients of whom one died with cytomegalovirus infection.

**Results**

We showed that samples from both groups have fixed variants at resistance sites and mixed infections. NGS sequencing also revealed non-fixed variants at resistance sites in most of the patients who died (6/9). A machine learning approach identified 10 genes with non-fixed variants in these patients. These genes formed a viral signature which discriminated patients with cytomegalovirus infection who died from those that survived with high accuracy (AUC=0.96). Lymphocyte numbers for a subset of patients showed no recovery post-transplant in the patients who died.

**Conclusions**

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

**ABBREVIATIONS:**

HCMV – Human Cytomegalovirus

Kbp – kilobase pairs

NGS – Next Generation Sequencing

WGS – Whole Genome Sequencing

SOT – Solid organ transplant

SCT - haematopoietic stem cell transplantation

PID – primary immunodeficiency

BMT – bone marrow transplant

GCV – ganciclovir

FOS – foscarnet

CDV – cidofovir

MV – minority variant

**INTRODUCTION**

Human cytomegalovirus (HCMV; human herpesvirus 5) is a member of the *Betaherpesvirinae* subfamily with a worldwide seroprevalence of between 18-100% [1,2]. HCMV is usually a benign viral infection in immunocompetent individuals, however, it is a significant cause of morbidity and mortality in immunosuppressed patients [3,4]. Therefore, strategies for prevention as well as treatment are of paramount importance for transplant clinical success. Several therapies exist for prophylaxis, pre-emptive therapy and/or treatment of HCMV [5]. Treatment with ganciclovir (GCV), foscarnet (FOS), cidofovir (CDV) or letermovir has improved outcomes [6–8], although late resistance often occurs [9]. Despite excellent outcomes for most haematopoietic stem cell transplants (SCT) and solid organ (SOT) transplant recipients, severe life-threatening HCMV disease can develop in approximately 20% [7] to 50% of cases [10]. Next-generation sequencing (NGS) has associated the presence of fixed drug mutations and mixed infections with poorer outcomes [11–14]. To further investigate the pathogenesis of life-threatening HCMV in immunocompromised patients we analysed the viral populations of 16 children with persistent HCMV viraemia. We also analysed a cohort (n=35) of immunocompromised adults with persistent HCMV.

**METHODS**

**Sample collection and ethics**

*Great Ormond Street Hospital samples:*

Whole blood samples were stored at Great Ormond Street Hospital for Children (GOSH) at -80C. These residual samples were collected as part of the standard clinical care at GOSH, and subsequently approved for research use through the UCL Partners Infection DNA Bank by the NRES Committee London Fulham (REC reference: 12/LO/1089) and West Midlands Black Country Research Ethics Committee (REC reference: 18/WM/0186). All samples were anonymised. Informed patient consent was not required.

*Royal Free London samples:*

Samples were collected as part of the Wellcome collaborative grant 204870/Z/16/Z UKRI). UCL17-0008 Analysis of Cytomegalovirus Pathogenesis in Solid Organ Transplant Patients approved by the NRES Committee London Queens Square Ethics Committee (REC reference 17/LO/0916).

*Sequencing:*

Nucleic acid was enriched using custom baits and sequenced as previously described [11,12,15] (Supplementary Materials).

**Data availability**

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

**Statistical analysis**

Bioinformatics processing and statistical analyses are described in the Supplementary Materials. The viral signature score model and data, along with a script for calculating scores for new samples, can be found in the GitHub repository: <https://github.com/ucl-pathgenomics/HCMV_ViralSignature>.

**RESULTS**

## **Patients’ characteristics**

We analysed 16 retrospectively identified paediatric patients from Great Ormond Street Hospital for Children (GOSH) with primary immunodeficiency syndromes (PIDs), SCTs or SOTs (Supplementary Database 1). All had HCMV viraemia persisting with ≤0.5 log reduction despite antiviral treatment for 21 days or longer, which has been defined as refractory [16]. No patients received prophylaxis against HCMV, although all SCTs received standard acyclovir prophylaxis against alpha-herpesviruses. Pre-emptive antiviral treatment for HCMV was initiated at first detection in the PIDs when viraemia exceeded 1000 IU/ml in the SCT recipients and 3000 IU/ml in the SOTs. First-line therapy was ganciclovir in the SOTs and PIDs and foscarnet in the SCT recipients. We stratified patients into two groups: a poor outcome group, defined as those who died with HCMV viraemia (n=8) and a good outcome group defined as patients who cleared their HCMV (n=8). We analysed an additional cohort of 35 adult SOT recipients with persistent HCMV viraemia. Patients underwent either liver or kidney transplant at the Royal Free Hospital. Pre-emptive antiviral treatment was started with valganciclovir and ganciclovir on the first positive (>200 ge/ml) and stopped on the second negative PCR (<200 ge/ml). Subsequent viraemia >3000 ge/ml was treated again with valganciclovir and ganciclovir. In this cohort, only one patient (liver recipient) died following persistent HCMV viraemia.

## **Sequencing data**

We analysed a total of 141 HCMV sequences: 59 samples from 16 immunocompromised children (1-9 samples per patient) collected over time and 82 samples from 35 immunocompromised adults. All samples selected for this study had an average sequencing depth of unique reads of ≥10 reads/nucleotide (nt), and ≥95% coverage of the strain Merlin genome [17]. The average sequencing depth in these two cohorts ranged from 10x to 1407x (after removing duplicates) (Supplementary Database 1).

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

To investigate the presence of multiple viral strains, we first calculated genome-wide within-host diversity (π) for each sample [12] (Figure 1A) and we reconstructed haplotypes for suspected mixed infections (Figure 1B and Supplementary Figure 1)[18]. We identified a total of 14 mixed infections in both cohorts. Taken together, mixed infections were not predictive of clinical outcome (26% of patients with good outcomes with multiple strains versus 33% with poor outcomes,$X^{2}$=0.19, p-value=.66).

## **Drug resistance mutations and minority variants**

We investigated mutations in the UL97 (serine/threonine protein kinase) and the UL54 (DNA polymerase) genes, which are the targets of the anti-HCMV drugs used here (GCV, FOS) [19].

In the paediatric cohort, two patients (P22 and P26) showed fixed mutations at resistance sites (i.e., present in most of the viral population) (Table 1). In the adult cohort, we identified fixed variants at resistance sites in 5 patients (R01-00014, H01-00017, H01-00016, H01-00012, H01-00003). Overall, we did not find any difference between patients based on the presence of fixed variants at resistance sites (5/42 patients, 11.9%, in the good outcome group and 3/9, 33.5%, in the poor outcome group, $X^{2}$=2.57, p-value=.19).

We then investigate the presence of minority variants (MVs) (i.e., present in the minority of the viral population). To take into account the presence of artefactual variability in sequence samples, only MVs occurring at a frequency of ≥2% [20] and minimum variant depth ≥5 reads were considered. In both cohorts, NGS sequencing revealed low frequency GCV and FOS resistance mutations in 6/9 of the poor outcome patients, two of whom also had mixed infections and 2/42 in the good prognosis group, none of whom had mixed infections (Table 1, $X^{2}$=21.47, p-value=.00001). Most variants occurred at frequencies <15% (median frequency 13.45) (Supplementary Figure 2). In all patients who died low frequency resistance mutations persisted in multiple longitudinal samples (when available), with the majority failing to rise to fixation (Supplementary Figure 2). In contrast, drug resistance MVs present in the two patients who did well (adult cohort) did rise to fixation in later samples (Supplementary Figure 2).

We did not identify any resistance mutations in our sequenced samples for P10 (n=6, including days 5, 11, 14, 18, and days 75 and 174 post-admission), but Sanger sequencing detected two resistance mutations: L501I in UL54 only on day 18 of treatment (day 43 post-admission) and G598D in UL97 only on day 81 post-admission (treatment day 56).

## **Minority variants in all viral genes**

To better investigate the dynamics of HCMV genome variation, we expanded our analysis to include MVs in the whole genome in single infections (Supplementary Figure 3). Interestingly, non-synonymous (NS) HCMV minority variants (MVs) were not confined to antiviral resistance sites but were distributed randomly across the HCMV genome in both the paediatric and adult cohorts (Supplementary Figure 4).

To investigate whether specific regions were enriched for NS MVs, we directly compared and ranked HCMV genes that discriminated patients who died from patients who survived using machine learning methods. We combined the two cohorts to increase statistical power. The gene selection process identified 10 genes (K score > 8, p-values < 0.005, adjusted p-value <0.5) (Supplementary Figure 5, Supplementary Table 1) that showed more within-host variability in the poor outcome group compared to the good outcome group. No gene showed the opposite trend.

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

We focussed on NS MVs as these gave better discrimination between poor and good outcome groups than non-synonymous and synonymous mutations combined, for all genes, bar UL11, UL7 and UL97 (Supplementary Table 1).

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

We assessed the power of our ten-genes viral signature to predict poor clinical outcomes in HCMV samples. Two models were employed: one considering the presence/absence of MVs in all 10 genes and another focussing on known resistance genes (UL54 and UL97). Generalised logistic models (glm) were used, and ROC curves with AUC (area under the curve) were employed for model evaluation. The full model achieved an AUC of 0.96, significantly outperforming the model with only resistance genes (p-value < 0.001, Anova, Figure 2A). Probability estimates generated by the ten-genes glm model indicated the likelihood of unfavourable clinical outcomes for each observation (Figure 2B).

Only 5 samples from 3 patients were misclassified by the full model (Supplementary Table 2, Figure 2B). In the adult cohort, one sample from patient H01-00017 who survived was classified as “poor outcome”. A second sample (45 days later) had however a probability of 0% of being poor outcome. This patient was one of only two who survived with multiple resistance mutations, one fixed and another that rose to fixation in the second sample. A second patient (R01-00014) from the adult cohort had one sample misclassified (1/4). This patient died and most samples showed high probabilities of poor outcome. Patient P16 from the paediatric cohort showed a more complex picture over time. This patient died and most samples (4/7) showed concordant probabilities > 50% of poor outcome. The samples with high probabilities were interspersed with samples showing low probabilities of poor outcome. All of these were taken before SCT in the first twenty days after admission. We found that these discrepancies were not due to lower viral loads or samples’ quality (Supplementary Table 2), but they probably reflect the complex HCMV dynamics and the need for repeated testing for accurate results.

We also assessed the predictive power of the signature including mixed infections. The full model including mixed infections had a high predictive power (AUC=0.91), albeit lower than the model with only single infections (AUC=0.96), likely due to the difficulty in assembly and calling minority variants where multiple strains are present (Supplementary Figure 6).

To validate whether our signature was specific for immune-compromised individuals following anti-viral treatment, we examined sequences (n=29, from amniotic fluid) from congenital HCMV infections publicly available (Supplementary Table 3). Although congenital HCMV infections had higher variability than samples from immunocompromised patients (Supplementary Figures 7 and 8), we found fewer MVs in the 10 genes of the viral signature than in samples from patients who died.

## **Viral signature over time**

To determine how early MVs in the ten sentinel genes can be used to predict a potentially poor outcome, we plotted the probability of being in the poor outcome group for three patients who died and had longitudinal samples (Figure 3). We also plotted longitudinal data for the two patients who recovered and had MVs at resistance sites. We did not have samples earlier during HCMV infection for patients 22 and R01-00014. However, samples taken at days 171 (from admission) and 91 (from transplant) respectively (62 and 109 days before death) were positive for the predictive signature. In patient P16 the signature was present as early as 9 days after SCT.

### **Biological significance of the MVs**

Most of the HCMV genome is under purifying selection [21], presenting on average a greater proportion of synonymous (S) changes compared to NS and stop codons. Surprisingly five of the ten genes in our viral signature (UL54, UL20, UL121, UL97 and UL74) reversed this trend with greater NS vs S MVs (Table 2). In the genes, NS variants tended to cluster closer together than expected by chance suggesting a functional role. In addition, most of the MVs (63%) mapped to HCMV variable loci identified comparing GenBank sequences. A higher overlap was observed for hypervariable genes (e.g. UL74 [22]) compared with drug resistance genes (e.g. UL54) (Table 2).

The clustering of variable residues is a feature of epitopes for which plasticity provides advantages in the face of host immunity. We identified known and predicted T cell epitopes (IEDB database) overlapping with amino-acid changes in patients who died in 8/10 genes (including the 5 genes with NS>S).

##

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

The finding that MVs are significantly more likely to occur in regions predicted to be immunogenic led us to explore how immunity might relate to the presence of these MVs. Lymphocyte counts were available from a subset of paediatric patients (n=7, P1, P2, P4, P10, P11, P22, P23) (Figure 4A). In patients P1 and P2 (who received SCTs) and patient P10 (who received gene therapy) lymphocyte counts recovered quickly after treatment (Figure 4A). In contrast, patients P4, P11, P22 and P23, who died, showed no recovery of lymphocyte count after SCT. Lymphocyte counts were persistently low in both groups just after SCT or gene therapy and started to increase at day 100 after transplant. Linear mixed effect modelling showed a significant difference in the counts over time (Figure 4B, p-value < .001) with significant differences in the final lymphocyte counts (good outcome median lymphocyte count: 8.34, 95%CI: 6.69- 8.34; poor outcome median lymphocyte count: 0.275, 95%CI: 0.14-1.10).

Analysing the SOT adult cohort separately (subset n=10 patients), as lymphocyte counts change with age, patient R01-00014 who died also showed persistently lower lymphocyte counts for months after receiving liver transplant as compared with the rest of the SOT cohort (Supplementary Figures 9 and 10) (last time point before death for R01-00014 was 0.22, the median in the rest of the SOT patients was 1.61, 95% 0.68-1.61).

**DISCUSSION**

Cytomegalovirus (HCMV) is the most common cause of infection following bone marrow and solid organ transplants [7,23]. The mechanisms by which HCMV infection influences transplant outcome are not known [7], but drug-resistant HCMV strains and infection with multiple strains have been associated with increased morbidity and mortality [11–14].

To investigate viral factors influencing transplant outcome, we sequenced longitudinal HCMV samples from 16 immuno-compromised children with persistent viraemia. Half of this cohort died of HCMV complications. We also sequenced samples from 35 adult SOT recipients where only one patient died following persistent levels of HCMV viraemia.

Multiple-strain infections are common in immune-compromised individuals; in our study, we identified a slightly higher percentage of mixed infections in patients who died (33% vs 26%), but the difference was not significant.

The use of antiviral drugs in the treatment of HCMV disease perturbs the viral population, selecting for drug-resistance variants. About 25% of the patients analysed in this study showed resistance mutations at various frequencies in the DNA polymerase UL54 and the protein kinase UL97, which are the major drug targets. Fixed mutations were present in patients independently of the outcome. In contrast, minority variants were almost exclusively present in samples from patients who died. Interestingly, low frequencies resistant variants detected in two patients who survived, quickly rose to fixation, whereas those detected in three patients who died persisted at low frequencies in the longitudinal samples.

Compared to traditional sequencing, next-generation sequencing can detect resistance mutations at low frequency at high resolution, enabling the detection of evolving virus populations in immunocompromised individuals selected under anti-viral treatment [11,17,24,25]. Thus, the finding of MVs at drug resistance sites should trigger repeat testing to better define the phenotype as well as to identify early resistance mutations that may become fixed and require treatment change.

These data and previous observations confirm that HCMV is highly stable at the consensus level in immunocompromised patients with very few substitutions observed over time in single-strain infections (0-25 substitutions) [12,17,26]. To further investigate the greater within-host viral variation in some patients, we used a machine-learning approach to attempt to discriminate between patients who died and those who survived. Using only samples from single infections, we identified the presence of MVs in one or more of 10 genes, including UL54 and UL97 as discriminatory between the two groups. Notwithstanding the opportunistic nature of the samples available, we were able to detect this signature on average 84 days before death and <100 days post-transplant and in all cases the signature was present in the first available sample. Interestingly, we identified the signature even in samples without resistance mutations from patients who died (P16, P23, P17). In addition, our signature seems to be distinctive for samples in immune-compromised patients as we could not find it in samples from congenital infections.

Only two genes of the signature were involved in drug resistance. Half of the genes included in the signature had a higher proportion of NS variants than expected by purifying selection [21], mapped to known HCMV variable genes and these loci clustered more than expected by chance. These results suggest positive selection, a hallmark of immune epitopes (Table 2) and indeed, in 7 out of 10 of the signature protein genes MVs mapped to known HCMV T cell epitopes. There might be several reasons why these variants remain at low frequency. Although variation at the consensus level is rare due to the proofreading activity of the viral DNA polymerase [27], one possibility is that low level variation in these epitopes occurs normally, but is cleared by functional T cell immunity. Variants are unlikely to confer increased fitness, rising to fixation only in circumstances where they enable evasion of prevailing immunity. In the absence of functional T cell immunity, as in the patients who died described here, we postulate that variants arising in epitopes can persist at low frequencies long enough to allow detection by NGS sequencing. In addition, there is evidence that GCV resistance mutations are not evenly distributed in different cell compartments [28,29] and the presence of low level virus subpopulations with antiviral resistance may represent virus confined in certain cell types.

Taken together the data hint at the possibility that dysregulated immunity contributes to the accumulation of MVs. Early studies showed that recovery of CD8+ T cells and CD4+ T cells is a positive predictor for prevention of mortality in HCMV-disease [30–32]. Restoration of HCMV-specific CTL response (class I MHC-restricted specific CD8+ CTL) may require an extended time after transplant in some patients, and such patients are at increased risk of developing severe HCMV disease. In our study, we were not able to obtain a measurement of T cell function, largely because the peripheral blood lymphocyte subset counts were too low for the assays used. Instead, we analysed lymphocyte counts as a proxy for lymphocyte function in a subset of 7 children for whom data was available. None of the four patients (three post-SCT and one with PID) who died had measurable lymphocyte counts and all harboured viral variants in the ten genes as described above. In contrast, three subjects with a good outcome (two post-SCT and one after gene therapy) for whom we had data showed good lymphocyte count recovery. Thus, the detection of persistent low-frequency HCMV mutations may be a biomarker of poor immune reconstitution and consequent poor outcome of HCMV infection.

Although fatal HCMV disease is less common in SOTs, it is interesting that patient R01-00014 (from the adult cohort) who died with disseminated HCMV showed a similar signature to the paediatric patients, suggesting that similar processes may underlie fatal HCMV disease irrespective of transplant type.

In this opportunistically collected sample set, we did not always have samples early on in HCMV infection. Notwithstanding, the viral signature was present, in all cases in the first sample tested. In all cases the signature was detected <100 days after transplant, i.e., before T cell recovery is expected, thus providing a potential early biomarker for failure of engraftment and poor outcome of HCMV infection. Since low-frequency resistance mutations which later rise to fixation can occur in the good prognosis group, repeated testing to demonstrate persistent minority variants is likely to increase specificity.

This study comprises initial observations drawn from a relatively small cohort, particularly within the poor outcome group, with samples from children collected opportunistically during routine clinical care at GOSH. Given the opportunistic nature of sample collection, these observations lack the structure of a pre-planned study. To validate and extend our findings it will be crucial to conduct prospective studies with a larger and more diverse patient population, incorporating longitudinal sampling. The importance of pre-planned sampling cannot be emphasized enough, as it enables the incorporation of technical and clinical replicates, enhancing the robustness and generalisability of our results. A further limitation is that the biological basis for these observations is not known although we speculate as to a possible explanation.

Despite the availability of effective antivirals, HCMV remains a serious infection, particularly in the context of immunocompromised individuals. Routine use of NGS in refractory patients could potentially detect significant resistance at earlier time points. At the same time, repeated detection of MVs may prove to be a useful biomarker for poor response to drug treatment alone and identify patients for whom non-pharmaceutical rescue therapies may be needed.

**FUNDING:**

This work was supported by the Wellcome Trust (Grant No. “204870/Z/16/Z” to CV, CA and CF). JB receives funding from the National Institute for Health and and Care Research (NIHR) UCL/UCLH Biomedical Research Centre.

**POTENTIAL CONFLICT OF INTEREST:**

The authors declare that they have no conflicts of interest related to the research in this manuscript.

**ACKNOWLEDGMENTS:**

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

**REFERENCES**

1. Cannon MJ, Schmid DS, Hyde TB. Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. Rev Med Virol. **2010**; 20(4):202–213.

2. Zuhair M, Smit GSA, Wallis G, et al. Estimation of the worldwide seroprevalence of cytomegalovirus: A systematic review and meta-analysis. Rev Med Virol. **2019**; 29(3):e2034.

3. Green ML, Leisenring W, Xie H, et al. Cytomegalovirus viral load and mortality after haemopoietic stem cell transplantation in the era of pre-emptive therapy: a retrospective cohort study. Lancet Haematol. **2016**; 3(3):e119–e127.

4. Ramanan P, Razonable RR. Cytomegalovirus Infections in Solid Organ Transplantation: A Review. Infect Chemother. The Korean Society of Infectious Diseases and Korean Society for Chemotherapy; **2013**; 45(3):260–271.

5. Kotton CN, Kumar D, Caliendo AM, et al. The Third International Consensus Guidelines on the Management of Cytomegalovirus in Solid-organ Transplantation. Transplantation. **2018**; 102(6):900–931.

6. Boeckh M, Ljungman P. How we treat cytomegalovirus in hematopoietic cell transplant recipients. Blood. **2009**; 113(23):5711–5719.

7. Kotton CN. Management of cytomegalovirus infection in solid organ transplantation. Nat Rev Nephrol. **2010**; 6(12):711–721.

8. Bruminhent J, Razonable RR. Advances in drug therapies for cytomegalovirus in transplantation: a focus on maribavir and letermovir. Expert Opin Orphan Drugs. Taylor & Francis; **2020**; 8(10):393–401.

9. Shmueli E, Or R, Shapira MY, et al. High Rate of Cytomegalovirus Drug Resistance Among Patients Receiving Preemptive Antiviral Treatment After Haploidentical Stem Cell Transplantation. J Infect Dis. **2014**; 209(4):557–561.

10. Beek MT van der, Marijt EW, Vossen AC, et al. Failure of Pre-Emptive Treatment of Cytomegalovirus Infections and Antiviral Resistance in Stem Cell Transplant Recipients. Antivir Ther. SAGE Publications Ltd STM; **2012**; 17(1):45–51.

11. Houldcroft CJ, Bryant JM, Depledge DP, et al. Detection of Low Frequency Multi-Drug Resistance and Novel Putative Maribavir Resistance in Immunocompromised Pediatric Patients with Cytomegalovirus. Front Microbiol [Internet]. Frontiers; **2016** [cited 2020 May 27]; 7. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2016.01317/full

12. Cudini J, Roy S, Houldcroft CJ, et al. Human cytomegalovirus haplotype reconstruction reveals high diversity due to superinfection and evidence of within-host recombination. Proc Natl Acad Sci. **2019**; 116(12):5693–5698.

13. Coaquette A, Bourgeois A, Dirand C, Varin A, Chen W, Herbein G. Mixed Cytomegalovirus Glycoprotein B Genotypes in Immunocompromised Patients. Clin Infect Dis. **2004**; 39(2):155–161.

14. Lisboa LF, Tong Y, Kumar D, et al. Analysis and clinical correlation of genetic variation in cytomegalovirus. Transpl Infect Dis. **2012**; 14(2):132–140.

15. Depledge DP, Palser AL, Watson SJ, et al. Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. PLOS ONE. Public Library of Science; **2011**; 6(11):e27805.

16. Chemaly RF, Chou S, Einsele H, et al. Definitions of Resistant and Refractory Cytomegalovirus Infection and Disease in Transplant Recipients for Use in Clinical Trials. Clin Infect Dis. **2019**; 68(8):1420–1426.

17. Suárez NM, Blyth E, Li K, et al. Whole-Genome Approach to Assessing Human Cytomegalovirus Dynamics in Transplant Patients Undergoing Antiviral Therapy. Front Cell Infect Microbiol [Internet]. Frontiers; **2020** [cited 2020 Aug 18]; 10. Available from: https://www.frontiersin.org/articles/10.3389/fcimb.2020.00267/full#h3

18. Venturini C, Pang J, Tamuri AU, et al. Haplotype assignment of longitudinal viral deep-sequencing data using co-variation of variant frequencies. Virus Evol [Internet]. Oxford University Press (OUP); **2022** [cited 2022 Oct 31]; . Available from: https://doi.org/10.1093/ve/veac093

19. Charles OJ, Venturini C, Breuer J. cmvdrg - An R package for Human Cytomegalovirus antiviral Drug Resistance Genotyping [Internet]. bioRxiv; 2020 [cited 2022 Apr 5]. p. 2020.05.15.097907. Available from: https://www.biorxiv.org/content/10.1101/2020.05.15.097907v1

20. Xu C, Nezami Ranjbar MR, Wu Z, DiCarlo J, Wang Y. Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller. BMC Genomics. **2017**; 18(1):5.

21. Lassalle F, Depledge DP, Reeves MB, et al. Islands of linkage in an ocean of pervasive recombination reveals two-speed evolution of human cytomegalovirus genomes. Virus Evol. **2016**; 2(1):vew017.

22. Suárez NM, Musonda KG, Escriva E, et al. Multiple-Strain Infections of Human Cytomegalovirus With High Genomic Diversity Are Common in Breast Milk From Human Immunodeficiency Virus–Infected Women in Zambia. J Infect Dis. **2019**; 220(5):792–801.

23. Hiwarkar P, Gaspar HB, Gilmour K, et al. Impact of viral reactivations in the era of pre-emptive antiviral drug therapy following allogeneic haematopoietic SCT in paediatric recipients. Bone Marrow Transplant. Nature Publishing Group; **2013**; 48(6):803–808.

24. Chou S, Ercolani RJ, Sahoo MK, Lefterova MI, Strasfeld LM, Pinsky BA. Improved Detection of Emerging Drug-Resistant Mutant Cytomegalovirus Subpopulations by Deep Sequencing. Antimicrob Agents Chemother. **2014**; 58(8):4697–4702.

25. Guermouche H, Burrel S, Mercier-Darty M, et al. Characterization of the dynamics of human cytomegalovirus resistance to antiviral drugs by ultra-deep sequencing. Antiviral Res. **2020**; 173:104647.

26. Hage E, Wilkie GS, Linnenweber-Held S, et al. Characterization of Human Cytomegalovirus Genome Diversity in Immunocompromised Hosts by Whole-Genome Sequencing Directly From Clinical Specimens. J Infect Dis. **2017**; 215(11):1673–1683.

27. Renzette N, Gibson L, Jensen JD, Kowalik TF. Human cytomegalovirus intrahost evolution—a new avenue for understanding and controlling herpesvirus infections. Curr Opin Virol. **2014**; 8:109–115.

28. Eckle T, Prix L, Jahn G, et al. Drug-resistant human cytomegalovirus infection in children after allogeneic stem cell transplantation may have different clinical outcomes. Blood. **2000**; 96(9):3286–3289.

29. Frange P, Boutolleau D, Leruez-Ville M, et al. Temporal and spatial compartmentalization of drug-resistant cytomegalovirus (CMV) in a child with CMV meningoencephalitis: Implications for sampling in molecular diagnosis. J Clin Microbiol. **2013**; 51(12):4266–4269.

30. Griffiths P, Reeves M. Pathogenesis of human cytomegalovirus in the immunocompromised host. Nat Rev Microbiol. Nature Publishing Group; **2021**; 19(12):759–773.

31. Reusser P, Riddell S, Meyers J, Greenberg P. Cytotoxic T-lymphocyte response to cytomegalovirus after human allogeneic bone marrow transplantation: pattern of recovery and correlation with cytomegalovirus infection and disease. Blood. **1991**; 78(5):1373–1380.

32. Quinnan GV, Kirmani N, Rook AH, et al. Cytotoxic T Cells in Cytomegalovirus Infection. N Engl J Med. Massachusetts Medical Society; **1982**; 307(1):7–13.

33. Pang J, Slyker JA, Roy S, et al. Mixed cytomegalovirus genotypes in HIV-positive mothers show compartmentalization and distinct patterns of transmission to infants. Stanley M, Akhmanova A, Ramchandar N, editors. eLife. eLife Sciences Publications, Ltd; **2020**; 9:e63199.

**TABLES**

**Table 1.** All detected drug resistance mutations (FC >=2)

|  |  |  |  |
| --- | --- | --- | --- |
| **Clinical outcome** | **Patient** | **Fixed**  | **Minority variants** |
| **UL54** | **UL97**  | **UL54** | **UL97** |
| Poor | P22 | K513N+, Q578L | M460I | E756D, Q578La, 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 | L545Sa | M460V |
| 3 patients (P16, P17 and P23) with poor clinical outcome did not show any resistance mutations  |
| Good | H01-00017 | N408K | M460I | N408Ka |  |
| H01-00016 |  | L595S |  |  |
| H01-00012 |  | C603W |  | C603Wa |
| H01-00003 |  | L595S |  |  |
| P10 | L501I\* | G598D\* |  |  |
| 37 patients with good clinical outcome did not show any resistance mutations |

\*Detected only by Sanger sequencing

+ Detected by Sanger sequencing and NGS

a Variants rising to fixation

**Table 2**. The table shows biological features of the ten HCMV genes under investigation in patients who died. The table shows the % of NS variants mapping to HCMV variable sites, the number of NS vs S MVs found, p-values indicating whether the NS MVs clustered significantly closer than by chance and known and predicted T cell epitopes from the IEDB database which MVs mapped to (in bold and underline the position of the MV in the epitope).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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**A**LK;L**E**NGVTHRF;NHGAGG**T**AAVSYQGA |
| UL20 | Uncharacterised | 92%  | 57-42 | 13-21 | p-value=3.6e-03 | MLG**IR**AMLVMLDYYW; SST**EGN**WSVTNLTES; MLL**PR**QYTL; FMDY**V**ILTP**L**AVLTC; |
| UL11 | Plays a role in the modulation of host immune response by modulating T cell function | 44%  | 34-392 stop codon | 20-17 | p-value=7.6e-04 | CYYVYV**TQ**NGTLPTT |
| UL8 | Membrane protein | 83%  | 47-47 | 23-35 | p-value=6.5e-01 | **S**SD**WV**TLGTSA**S**LL**R** |
| 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**T**AALW; ISL**V**TPLTINATLRL; SCTHPYVISL**V**TPLT |
| UL75 | Envelope glycoprotein gH | 100%  | 25 - 63 | 8-18 | p-value=9e-04 | FPDATV**P**ATV; K**A**QLNRHSYLKDSDFLDAA; RQTEKHELLVLVKK**A**QLNRH; HELLVLVKK**A**QL; YLLSHLPSQRYGADAASEALD**P**HAFHLLLNTYGRPIRFLRENTTQC; A**A**SE**A**LD**P**HAFHLLLNTYGR; LD**K**AFHLLL; YL**L**SHL**P**SQRYGA**D**A**A**SE**A**LDPHAFHLLLNTYGRPIRFLRENTTQC |
| UL7 | CEACAM1-like protein; plays a role in modulating the host immune response  | 78%  | 20-33 | 13-20 | p-value=3.5e-02 | STPYVGLS**LS**CAANQ |
| 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 | LLFLD**E**IRNFSL**RS**P; TMRK**L**KRKQALVKEQ; SFY**L**VNAMSRNLFRV |

**FIGURES LEGEND**

**Figure 1.** **A)** 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. Mixed infections are labelled with “M”. **B)** 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 [33]. **C)** Summary heatmap showing patients with mixed infection, patients with known fixed or minority variants at resistance sites and cohort information (green for the adult and violet for the paediatric cohort).

**Figure 2.** **A)** ROC curves with confidence intervals (95%) for two predictive models discriminating between samples from patients who died and survivors. AUC for the full model (including MVs in the 10 candidate genes) was 0.96 (red ROC curve). AUC for the drug resistance genes model (including genes UL54 and UL97) was 0.65 (green ROC curve). **B**) Estimated probabilities for each sample in the two groups (red for survivors, turquoise for patients who died) to be classified as a patient from the poor outcome group. Arrows and circle indicate patients with at least a sample which was misclassified by the model.

**Figure 3.** Viraemia, anti-viral therapy and transplant, variants at resistance sites and viral signature probability in patients with longitudinal samples and low-level resistance mutations, including patient 16, 22 and R01-00014 from the poor outcome group and H01-00012 and H01-00017 from the good outcome group.  Dots indicate samples sequenced and are coloured based on viral signature probabilities (from yellow, good outcome, to black, poor outcome). Black rectangles indicate anti-viral treatment, and the blue circles shows the time of transplant (SCT for patient 16, liver for patient R01-00014 and thymus for patient 22, kidney for patients H01-00012 and H01-00017).

**Figure 4. A)** Lymphocyte count (per microliter of blood) overtime in a subset of GOSH patients. Time of SCT or gene therapy is shown in red. In orange we indicate the healthy lymphocyte count range for children (3-13). Patients with poor outcomes are indicated with a red X. **B)** Trend lines (smoothed local regression line using loess) for lymphocyte count for and poor outcome groups after SCT or gene therapy (or after admission for patients who did not have SCT/gene therapy). The grey area represents 95% CI.