Genomic and geographical structure in Human Cytomegalovirus

Oscar J. Charles[1,+], Cristina Venturini[1,+,*], Soren Gantt[2], Claire Atkinson[3], Paul Griffiths[3], Richard A. Goldstein[4] and Judith Breuer[1,4]

1. Department of Infection, Immunity and Inflammation, UCL Great Ormond Street Institute of Child Health, London, UK
2. Research Centre of the Sainte-Justine University Hospital and Department of Microbiology, Infectious Diseases and Immunology, University of Montréal, Canada
3. Division of Infection and Immunity, Institute for Immunity and Transplantation, University College London, London, UK
4. Division of Infection and Immunity, University College London, The Cruciform Building, Gower St, London, UK
5. Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK

[+]These authors contributed equally to this work.

[*] Corresponding author:

Cristina Venturini

c.venturini@ucl.ac.uk

Classification

Biological Sciences

Evolution

Keywords

Human cytomegalovirus, genotyping, hypervariability, genomics, Hidden Markov Models, phylogeography, molecular population genetics

Abstract

Human cytomegalovirus (CMV) has infected humans since the origin of our species and currently infects most of the world's population. Variability between CMV genomes is the highest of any human herpesvirus, yet large portions of the genome are conserved. Here we show that the genome encodes 74 regions of relatively high variability each with 2-8 alleles. We then identified two patterns in the CMV genome. Conserved parts of the genome and a minority (32) of variable regions show geographic population structure with evidence for African or European clustering, although hybrid strains are present. We find no evidence that geographic segregation has been driven by host immune pressure affecting known antigenic sites. Forty-two variable regions show no geographical structure, with similar allele distributions across different continental populations. These "non-geographical" regions are significantly enriched for genes encoding immunomodulatory functions suggesting a core functional importance. We hypothesise that at least two CMV founder populations account for the geographical differences that are largely seen in the conserved portions of the genome, although the timing of separation and direction of spread between the two is not clear. In contrast, the similar allele frequencies among 42 variable regions of the genome, irrespective of geographical origin, is indicative of a second evolutionary process, namely balancing selection that may preserve properties critical to CMV biological function. Given that genetic differences between CMV viruses are postulated to alter immunogenicity and potentially function, understanding these two evolutionary processes could contribute important information for the development of globally effective vaccines and the identification of novel drug targets.

Significance statement

CMV genome diversity is higher than other human herpesvirus, and recombination is pervasive. Here, using Hidden Markov modelling, we describe 74 multi-allelic regions, with the remaining 86% of the genome showing lower variability, albeit with single nucleotide polymorphisms. We demonstrate for the first time that CMV diversity is influenced by two distinct evolutionary forces. A founder effect results in geographical segregation affecting the regions of low variability and 32 variable regions. In contrast, the 42 remaining regions, which are enriched for immunomodulatory functions, show so-called balancing selection, resulting in maintenance of equal allele frequencies

irrespective of geography. These new insights into CMV evolution are likely to provide insights in virus biology and inform the development of drugs and global vaccines.

Introduction

Human cytomegalovirus (CMV) is a member of the *Betaherpesvirinae* that infects circa 66% to 90% of adults in any given country (1). Like all human herpesviruses, CMV is a linear double-stranded DNA virus that causes lifelong latent infection by establishing latency in long-lived cell populations, and periodically reactivates resulting in lytic viral replication (2, 3). CMV causes significant burden of disease in those with compromised immune systems (4), and is also the most common infectious cause of congenital disability worldwide (5). Because of this, developing a vaccine is a high public health priority (6).

At approximately 236 kb, CMV has the largest genome of all human herpesviruses (7) and the highest level of genetic diversity of all the known human herpesviruses (8, 9). The virus is known to readily undergo recombination (10, 11), and co-infection is frequently observed, especially in individuals with weakened immune systems (12). Most of the observed CMV diversity occurs in discrete hypervariable regions where sequences cluster into genotypes also known as alleles (13, 14). In some hypervariable genes, e.g. UL55 (glycoprotein B, gB), alleles have been defined (15) and used alone or in combination with other multi-allelic regions to genotype CMV and identify mixed infections (16, 17). However, attempts to correlate individual alleles with transmission and pathogenesis have so far been unclear or contradictory (14, 18–24).

Despite the efforts to define CMV diversity, our understanding of the evolutionary history that led to its present variation and whether this variability follows a global pattern is limited. CMV, like other herpesviruses, shows remarkable species-specificity, which results from long-term co-evolution and adaptation to the host (25, 26). In contrast, unlike other herpesviruses, human CMV (HCMV) whole genomes from clinical samples show little evidence of geographical or other population structure, with the exception of two Asian genomes that have been shown to cluster phylogenetically (10, 11). This, together with the absence of data from ancient CMV genomes, makes for uncertainty as to when and how current HCMV diversity evolved.

To better understand CMV genomic structure and how it has evolved, we employed Hidden Markov Model (HMM) clustering to delineate the hypervariable and conserved regions across a global and diverse dataset of published and unpublished CMV genomes, which together represent unrelated clinical and low-passage strains (27–29). HMM is able to determine the number of sequence clusters (i.e. alleles) that best explain the diversity across CMV genomes and to identify regions where multiple alleles are present. Using the outputs of the model we describe precise co-ordinates for regions of multi-allelic variability, some of which are novel. We also show, for the first time, that CMV genomes do display geographic population structure and that this is particularly clear within the relatively conserved monoallelic regions of the genomes. We highlight examples of where our approach can help to provide insights for further research into questions of viral pathogenesis and ancient evolution.

Results

CMV diversity and population structure is determined by 74 discrete multi-allelic regions

We compiled a set of 259 CMV whole genome sequences which had been collected worldwide (Supplementary figure 1, full details are shown in Dataset S1). 233 sequences were retrieved from GenBank (30) with available metadata for 106. Of these 106, 35 were from patients who were immunocompromised through HIV, organ, or bone marrow transplant and 71 were from immune-competent individuals (of whom 60 were congenitally infected babies). Short read data were available for 17 samples allowing us to check for mixed infections. Nine samples contained only a single CMV strain, of which five were from HIV-positive and CMV-positive mothers and four were immuno-competent children (primary infection or sibling). Eight samples showed evidence of multiple strains for which we reconstructed the haplotypes (17 sequences in total) with validated methods (8, 31). All eight samples came from HIV-positive mothers with CMV infection (8, 32).

Most genomes were derived from virus obtained in Europe and the Middle East (n=216, including Belgium, Czech Republic, France, Germany, Greece, Italy, Netherlands, United Kingdom, and Israel). As Israeli genomes appeared as if European in analyses, and the migratory history of the country is complex, we have labelled Israeli sampled genomes as European for simplicity. 30 sequences (including the reconstructed haplotypes) were from samples collected in Africa (Zambia, Kenya and Uganda) (17) (32–34). The remaining sequences were collected in other parts of the world: 11 from the United States (America); 2 from Asia (China and South Korea) and 1 from Australia (Oceania).

To characterise CMV genomic diversity we first aligned the genomes and then calculated heterozygosity along a sliding window of 50 base pairs (bp) (Track1 in red, Figure 1). This revealed that while most of the genome is highly conserved there are regions of significant nucleotide diversity, some of which have previously been described (35, 36). Aligning HCMV genomes in these hypervariable regions is challenging and leads to numerous gaps in alignments that are often ignored in phylogenetic and genetic distance calculations. To deal with the complexity of CMV genomes alignment in these regions, we developed a sequence clustering method based on Hidden Markov Model called “HMMcluster”. This approach is unaffected by gaps and it groups together sequences based on the statistical likelihood that they came from the same underlying source. The genome-wide implementation has two steps: in the first (1) a fixed window is defined (in this case 200bp, chosen as it returned regions no smaller than 20bp) and the model calculates the minimum number of sequence clusters (alleles) that best explain the diversity within the window (37). The second step (2) concatenates contiguous windows of sequence variation, and it refines the coordinates for a “variable region”. Variable regions with multiple alleles are defined as “multi-allelic”.

From this we identified 74 discrete variable regions for which the model provided statistical support for the presence of more than one allele (Track 3 in blue - Figure 1, Table S1, Table S2). These multi-allelic regions range in size from 26 to 4760 nucleotides, encompassing 14% of the genome (Figure 1, Table S1). The remaining 86% of the genome was found to be highly conserved with no statistical support for multiple alleles. Whilst multi-allelic regions generally occur in regions of high nucleotide diversity, some, generally smaller in size and with fewer segregating Single Nucleotide Polymorphisms (SNPs), were found in regions of otherwise lower nucleotide diversity. The 74 multi-allelic regions were not constrained to known coding regions and encompass the previously documented 12 hypervariable genes (7). Fifty-two of the regions were each contained entirely within a single gene, 17 crossed gene boundaries by either spanning the gap between two genes or extending beyond a gene terminus into non-coding regions, and 5 regions were entirely in non-coding or otherwise unassigned portions of the genome.

To characterise the evolutionary relationships between the alleles within each multi-allelic region we constructed maximum likelihood phylogenetic trees for each of the multi-allelic regions, excluding 10 with evidence of recombination breakpoints (Table 1 and Supplementary figure 2). For the remaining 64/74 we observed that multi-allelic regions consisted of well separated clades representing HMMcluster supported alleles, whereas reconstructed phylogenies of similarly sized conserved regions tended to show poorly resolved clades with low bootstrap support for key central nodes (Figure 2). Genetic variation in multi-allelic regions were as much as an order of magnitude greater than comparably sized conserved genome portions (Figure 2). Well separated clades with restricted recombination have been identified previously in CMV hypervariable regions and are thought to represent an inability of homologous strands to anneal, supporting the development of population structure within those loci (11).

European and African whole genomes display geographical population structure

To examine the relationship between CMV strains we first analysed whole genomes by the geographical region (continent) in which they were sampled. Geographical segregation of genomes is well described for other herpesviruses (38–41). Because CMV is known to be highly recombinant such that accurately reconstructing phylogenetic relationships and tree topology become problematic, we initially used Multi Dimensions Scaling (MDS) to analyse the genomes (Figure 3) as dimensionality reduction techniques simply represent "closeness" of sequences and have been able to derive representations of genetic data resembling their geographic ancestry (42, 43). The resulting clustering pattern showed geographic segregation of whole genomes in the second most important dimension (component 2, Supplementary Figure 3) with those sampled in Africa clustering away from most sequences sampled in Europe. Two sequences sampled in Asia were located amongst the European sequences (Figure 3). MDS components 1, 3 and 4 which represent large variance in the data were uninformative with respect to geographical segregation of whole genomes.

This geographical split was corroborated by a phylogenetic network tree, which allows representations of shared recombination events as network splits (Supplementary Figure 4). Sequences sampled in the Americas (all from the USA) were distributed throughout the plot. The single sequence from Oceania (Australia) appeared to resemble sequences sampled in Europe. A minority of sequences sampled in Europe clustered with sequences sampled in Africa, but not vice versa (Figure 3). To examine this in more detail, we made use of a separate set of CMV genomes from a cohort of CMV seropositive solid organ transplant recipients (n=11, Dataset S2) with available self-reported ethnicity data, all of whom were sampled in the UK (44). Performing MDS of these sequences along with the European, African, and Asian sampled CMV showed that CMV from transplant recipients of self-reported African and Afro-Caribbean origin but sampled in the UK clustered predominantly with strains sampled in Africa, and separately from UK patients of non-African origin (Supplementary Figure 5). This suggests that sequences cluster by host ancestry rather than simply by sampling location. The few European sampled strains that clustered with the bulk of African samples sequences are therefore likely to be from individuals of African ancestry. Our findings contrast with previously published results that failed to identify CMV population structure related to geographical origin in partial genomes (45, 46) or a subset of highly heterogenous whole genomes (11), although the two published Asian sequences have been noted as tightly clustering in an analysis of whole genomes (10).

Conserved and multi-allelic genomic regions show distinct patterns of phylogeography

Because of potential differences in evolutionary histories between conserved and multi-allelic regions we next analysed these separately for geographic structure. MDS of the concatenated conserved genomes (concatenome) showed more distinct separation between viruses of African and

European origin than the whole genome sequences, as well as clearer segregation of the Asian CMV viruses (Figure 3; Supplementary Figures 4 and 5). In contrast, multi-allelic regions when concatenated appeared to show minimal geographic clustering. We quantified the differences between African and European populations in conserved regions using Fixation Index (Fst) which compares diversity within and between different populations. As Fst can be biased if sample sizes vary between populations (47), we randomly chose 30 African-sampled and 30 European-sampled sequences. The continent labels were then randomly scrambled to generate a null hypothesis Fst and both steps were repeated 10,000 times to obtain true and null Fst distributions (Supplementary Figure 6).

The results showed that the conserved regions in the CMV genome encode clear geographic (continental) differences, with a mean Fst of 0.21 (a 423% increase on the mean null Fst). Fst of the concatenated multi-allelic regions was relatively weaker at 0.097, which is only a 194% increase on the mean null Fst. While both Fst values were significantly different to their respective null distributions (both Mann Whitney test p-values < 0.0005.), geographic signal appears to be more enriched (423% vs 194%) for the conserved regions of the CMV genome.

Admixture model-based estimation of ancestry supports continental population structure

To further examine the segregation of conserved genomic sequences by continent and to use a model based approach to complement the visual MDS and network phylogeny, we undertook an admixture analysis which attempts to infer the ancestral lineages and the contributions from each that gave rise to a set of modern sequences (Figure 4) (48). Admixture analyses, like dimensionality reduction, can also be skewed by large sample sizes differences between groups and by heterologous sampling methodology between groups (49). To account for the former, we randomly subsampled to generate more proportionate sample sizes per continent (30 European, 30 African and both Asian sequences), then calculated the Cross Validation Error (CVE) for K=1 through 10. This was repeated for 1000 random sample draws.

Two lineages (K=2), one African and one Eurasian were found to be the consensus result from the 10 lowest error models (Figure 4A and Supplementary figure 7A). CMV sequences sampled in Africa from individuals of African ancestry, were clearly identified by the model as being part of this African lineage, with no hint of admixing and vice versa for white Europeans. Few sequences sampled in Europe were assigned strongly to the African lineage but in the MDS these are likely explained as being sampled from ethnic African individuals in Europe (Figure 3A, 4B; Supplementary Figure 5). Many sequences also appeared as admixes and correspond well with those sequences that lie between the clusters of African and European in the MDS.

For a minority of subsampling replicates there was a random skew toward sampling more Asian-like CMV sequences and in three replicates the optimal model was K=3 clusters (Supplementary figure 7). Although the current data is best explained by two ancestral genomes (K=2), the few Asian sequences (n=2), although clustering together will not change the admixture result for most replicates (49). Notwithstanding, there was also clustering of two Asian CMV conserved concatenomes from UK patients of self-reported Asian ethnicity (Supplementary Figure 5 and Dataset S2) with the two Asian-origin GenBank sequences making it likely that, with more genomes three ancestral sequences (K=3) may turn out to better represent the data.

To overcome the complexity of sample origin not necessarily reflecting virus ancestry (i.e. human migration) and to allow better delineation of population differences between African and European CMV lineages, we limited further analyses to "archetypal" sequences from each continent. Archetypal sequences were defined as those with >90% admixture proportion to a single ancestral cluster using the conserved concatenome data, in the lowest error k=3 model. Using this definition, we identified 42 strains (12 of which were sampled in Europe) of archetypal African ancestry and 129 of archetypal European viral ancestry, which includes the reference strain Merlin. Four CMV sequences were identified as of archetypal Asian ancestry (2 of which were sampled in Europe from subjects with self-reported Asian ancestry).

Geographic signal is consistent across the conserved regions and sequences between continental clusters have random patterns ruling out recent recombination

To understand how the geographic signal is ordered across the viral genome we identified the loci within the conserved concatenome with high (>0.5) Fst values, i.e. which showed the most distinction between African and European strains. To provide a visual representation of the continent of origin for each strain, we coloured high Fst value African nucleotides present at consensus (>50%) in archetypal African strains red and high Fst value European nucleotides present at consensus (>50%) in archetypal European strains, blue. Sites which were not present at consensus in ether archetypal African or European strains were coloured (black) (Supplementary Figure 8). In the archetypal African or European subset, sequences were overwhelmingly represented by the base most common to their continent. Moreover, where an African consensus base appears in an archetypal European strain, or vice versa this appeared to be largely random. Sequences not archetypal of these three continents had complex, although apparently non-random patterns of admixture.

Multi-allelic genomic regions represent a mixture of geographic relationships

We next examined the apparent but weaker geographical segregation of the 74 multi-allelic regions between African and European populations. As the multi-allelic regions each contain variable numbers of alleles and are of different lengths, we considered each of the 74 regions separately. We performed chi squared tests with allele distributions to evaluate whether multi-allelic regions segregated with archetypal African or archetypal European strains. The results revealed strong geographic distribution of alleles for 32 of the 74 regions (Table S1, we calculated false discovery rate, FDR, using the Benjamini-Hochberg procedure and we set the FDR threshold at 0.05) (50), while five showed more moderate geographic differences (0.05 < FDR < 0.3; Table S1) with the rest showing no evidence of geographical segregation. Colouring African dominant alleles red and European-dominant alleles blue (as defined above) for the 32 multi-allelic regions that segregated geographically we observe a similar pattern to that seen for the conserved regions, with African strains largely red and European strains largely blue (Supplementary Figure 8, C and D).

Host Immune mediated selection is not obviously responsible for driving geographic segregation of African and European strains

To assess whether the geographic population structure observed in conserved regions of the CMV genome was due to differences in the selective pressures exerted by different host populations, we looked at the 440 most geographically informative sites (Fst > 0.5) within the archetypal subsets of European and African strains. We asked how many of these sites resulted in a different continental consensus amino acid between African and European sequences and, if so, whether the change occurred within known B and T cell epitopes, as recorded in the Immune Epitope DataBase (IEDB) (51). Only 15% (64) of the 440 sites encoded nonsynonymous changes. Of these, 16% (10 of 64) lay within known epitopes, compared with 14% (52 of 376) of synonymous sites, providing little evidence that geographic population structure in CMV is driven by continentally unique host immune pressure.

Alleles in immunomodulatory genes tend to maintain similar diversity across continents

We next tested whether certain gene functions were over-represented in three classes of genomic regions: conserved, geographically segregating multi-allelic, and non-geographically segregating multi-allelic regions. To do this we identified the genes lying within each region class and annotated their function, using three predefined functional groups from a published gene-ontology [Latency, Tropism and Immunomodulation] (52). We then determined whether key functional groups were significantly over/under-represented by region class (Supplementary Figure 9). We found the 74 multi-allelic regions together were significantly enriched for genes encoding immunomodulatory functions with the proportion of genes with immunomodulatory function of 0.371 (37.1%) where for all genes this proportion is 0.206 (FDR= 0.0096). This enrichment was clearer still if only the 42 multi-allelic regions with no evidence of geographical segregation were considered (39.1%, FDR=0.014). No other comparison was significant.

Geographic and multi-allelic genetic differences may impact biological function

One interesting application of studying multi-allelic regions is the possible association between different alleles, and function. To illustrate this, we investigated glycoprotein B (gB, UL55) as an example. The reference strain Towne has been extensively used to develop vaccines and to study CMV immunogenicity and function (53) and its gB protein variant has been the basis for many vaccine candidates (54). From the admixture analysis we identified Towne as being 79% African, while two other common reference strains AD169 and Merlin were 75% and 99% European respectively. gB (UL55) sequence is composed of 3 separated multi-allelic regions (22, 23 and 24 in Table S1) which are linked (Monte Carlo chi square p<0.001) in 12 possible combinations or haplotypes. In addition to showing geographically informative genetic differences from European strains in its conserved regions, Towne gB also differs in the sequence of its multi-allelic regions sharing the same 22,23,24 (UL55) haplotype as only 4% of the 259 CMV genomes sequenced here (Supplementary Figure 10). By contrast, the Merlin multi-allelic region 22,23,24 (UL55) haplotype is shared with 32% and AD169 with 20% of the 259 genomes. Genetic differences in gB (UL55) have been mooted to underlie observed differences in cross-strain neutralisation by antibodies raised against one strain (55, 56). In general, immunotherapies and drug development targeting CMV that rely on alleles that differ across geographic isolates, may now require further investigation as to whether treatment effect will be advantageous to only certain human populations.

Discussion

We have characterised the variability present in whole CMV genome sequences, including several known to be of African origin identifying both known hypervariable regions and over 40 which have not previously identified (10, 11, 13, 16, 57). Altogether we describe 74 hypervariable regions comprising 14% of the genome, all of which are multi-allelic. The remaining 86% of the genome is monoallelic and at least ten times less variable than the multi-allelic regions. While previous reports have identified around 30 multi-allelic hypervariable regions, they identified them by the gene in which they were located, despite in many cases, much of the gene concerned not being hypervariable and thus subject to different constraints on recombination and diversity. In contrast, our data precisely delineates the nucleotide coordinates of variable and conserved regions, using an unbiassed and consistent assignment model. This has allowed the identification of 17 multi-allelic regions that cross gene boundaries and 5 are entirely in non-coding or in otherwise unassigned. For each of the 74 regions, our statistical approach also returns the number of alleles that most parsimoniously fit the data, leading to between 2 and 8 alleles per region (Table S1). In some cases, for example UL146 and UL144 , the number of alleles we identify differs from previous numbers reported (58, 59). However, previous estimates were determined largely by visual inspection of whole-gene phylogenetic trees, a process which resulted in differences in reported allele numbers not only from us but between other authors (for example for UL55), with, on occasion some alleles not being reliably distinguishable (16, 60–62). The objective mathematical approach we have adopted provides clear and reproducible multi-allelic region boundaries and allele numbers, properties which will have advantages for standardising genotyping nomenclature.

In contrast to previous reports (10, 11) we observe clear geographical segregation of CMV genomes with evidence for African, European, mixed and possibly Asian genotypes. From an evolutionary viewpoint, geographical segregation of genomes is well described for other human herpesviruses including herpes simplex virus (HSV-1), varicella zoster virus (VZV) and Epstein Barr virus (EBV) (38–41, 63) and is therefore not surprising for CMV. Our data suggest that most geographically informative SNPs in CMV are in the monoallelic genomic portions and like the rest of the genome are under purifying selection (10). Unlike EBV, in which host immune selection drives local adaptation of virus to different human host populations, shaping the pattern of genetic diversity (64), we find no evidence that selection within immunogenic regions of the genome are a dominant driver of the observed genetic geographical differences. Instead, we postulate that genetic drift and bottleneck events such as founder viruses are plausible explanations for the population structure observed in CMV. If this is the case, there remains some difficulty in establishing the direction and date of split for European and African CMV populations, due to the low association between sampling date and distance from phylogenetic tree root and a mutation rate that has proven difficult to determine for double-stranded DNA viruses (65). The difficulty in determining mutations rates is likely to be, in part, a result of CMV's longstanding free recombination within these geographically isolated pockets. As 13 multi-allelic regions were found to contain alleles unique to Europe while the opposite was not seen for African viruses, this could be taken as evidence supporting a European origin of CMV, where Africa has restricted diversity. However, this is likely more simply explained as an artefact of the differences in size and sampling heterogeneity of our available genomes.

The finding of African-clustering CMV strains in patients self-reporting as being of Afro-Caribbean ethnicity, many of whom presumably have not lived in Africa can potentially be explained by early acquisition of CMV from family members and by assortative mating of racial groups. A similar effect has been observed for VZV in subjects of Afro-Caribbean origin growing up in the UK (41). In African countries most children are CMV positive by their first birthday (33, 66, 67). Early Infection, most likely acquired from maternal or sibling transmission (68) may explain why subjects of African origin living in Europe test positive for strains that cluster with known African strains. This would date the split of African and European CMV strains to at least 500 years ago, the time at which the first African slaves were transported to Europe (69). Although the actual separation is likely to be much older given even the highest estimates of CMV's mutation rate (70), or borrowing from rates presumably more accurately estimated for a similar virus, HSV1, for which there are ancient genomes available (71). With greater mixing of populations and the pervasive genome-wide recombination that occurs in CMV, we see evidence for increasing numbers of hybrid strains including some of the reference strains for example Towne and AD159 (Figure 4). This serves to further muddy insights into the phylogeographic origins of currently circulating strains of CMV and elucidating these undoubtedly requires additional and more granular worldwide sampling, as well as the inclusion of ancient CMV genetic material if these can be found.

While in 32 of the multi-allelic regions, the alleles, like the conserved regions, are different or differently distributed between geographical regions, most notably countries with predominantly European or African populations, from which most strains originate, the majority (42) show no evidence of geographical distribution, but instead appear to maintain the full allele palette in these genome portions across both continents and at similar frequencies. This effect has previously been observed in RNA viruses (72, 73). Initial studies into CMV virion envelope complexes, linking variants to function, have reported that allele differences can modulate virus cell tropism (74, 75). Similar examples from hepatitis C virus and human immunodeficiency virus (HIV) have shown genotypic differences to affect viral compartmentalisation (76), while certain hepatitis B virus genotypes appear to be associated with chronic infection (77). From our analyses, the non-geographically segregated multi-allelic regions were significantly enriched for genes encoding immunomodulatory functions. For example, region 6 which shows no geographical segregation encodes a portion of the non-recombinant haplotype RL11D block, RL11, RL12, RL13, UL1, UL2 and UL4, all of which are proven or predicted to be virion membrane glycoproteins (52). In addition, RL11D (region 6) variability has been suggested to be critical for the adaptation of CMV to different primate species (11). Of interest, many of the variable multi-allelic regions correspond to regions previously identified as being in local linkage disequilibrium and thus not affected by the pervasive recombination occurring throughout the more conserved regions of the genome (11). Taken together, our data strongly support the likelihood that CMV genome is the result of two distinct evolutionary forces, genetic drift occurring in segregated viral populations and so called frequency-dependent balancing selection (78), a form of adaptation that maintains pre-existing diversity in the face of genetic drift.

This granularity of CMV genome analysis allows deeper insights into how genome might be related to function. Following Towne gB + MF59 adjuvant vaccination, antibody titres to the antigenic AD2 region have been shown to correlate with better protection against post renal transplant CMV viremia (54–56). Baraniak and colleagues also showed that only ~50% of individuals vaccinated with Towne gB had detectable AD2 antibody response against gB peptides in an assay derived from the AD169 laboratory strain. The AD169 gB AD2 allele (region 24, UL55) differs from that found in Towne gB (Supplementary figure 8). Since AD2 antibodies are not broadly reactive (54), there is some question about what the Towne gB vaccine antibodies are recognising within the AD169 gB AD2 peptides. Multi-allelic region 24 (gB/UL55 AD2) also segregates differently between African (Towne-like) and European-like (AD169 and Merlin) viral populations. This together with the finding that the Towne gB conserved region carries predominantly African-segregating SNPS raises the possibility that a vaccine based on Towne gB sequence might not confer cross-protective immunity against European strains.

Notably, the CMV pentamer complex, which is being developed as part of an alternative multi-antigenic vaccine using the Merlin strain, contains no multi-allelic regions and may therefore be more tractable than gB (79). However, geographically related differences are still present and potentially need evaluation. For example, Q35K which segregates with African strains and L40P which segregates with European strains are both present within a known B cell neutralising epitope in the pentameric complex UL130 protein (IEDB ID: 142031).

These analyses are subject to limitations, the clearest of which is the potential biases related to the available sequences and the samples from which they were derived. First, the representation of genomes was heavily skewed towards Europe. Second, the European samples were typically collected from unrelated patients for clinical purposes, whereas African samples were obtained from study participants in southern and eastern Africa (Zambia, Uganda, and Kenya), many of whom were HIV co-infected. African-clustering sequences that were sampled in Europe likely reflect divergent sampling of unrelated individuals and may to some extent mitigate the bias of African strains. However, we had only two Asian strains collected in the early 2000s and no strains from south America or many other parts of the world. Until these gaps are closed our conclusions must remain incomplete. Seventeen African genomes were reconstructed from samples containing mixed CMV infections where generating consensus genomes has typically been challenging. However, HaRold, the program we used to reconstruct haplotypes (31) performs with high accuracy in validation exercises using simulated and real mixtures of CMV genomes containing known sequences. Lack of homology between some alleles within a multi-allelic region could be a limitation on constructing alignments for HMM. However, even for the most divergent alleles, the bordering 5’ and 3’ sequences are identical, a factor that mitigates this potential constraint. Finally, when considering B and T cell CMV epitopes, we are limited by epitopes included in the IEDB database which are largely generated for European strains. However, since most of the viruses analysed here were European, the conclusions that most nonsynonymous differences from African strains do not lie within epitopes is likely to be true.

Conclusion

Our findings provide several new insights into the genomic landscape of CMV. First, we identify and precisely delineate 74 discrete variable regions which consist of multiple alleles, showing that the rest of the genome (86%) is monoallelic. We identify for the first time, that CMV genomic evolution is shaped by two distinct processes: likely genetic drift occurring within geographically distinct populations and balancing selection which counteracts genetic drift to maintain similar diversity in variable multi-allelic regions irrespective of geographical location. We identify that variable regions under balancing selection are enriched for key CMV properties, highlighting that better characterisation of diversity in these regions is likely to be important for understanding CMV biology and control. Our results provide a genomic roadmap to enable studies of how variation across the CMV genome interacts to cause clinical disease. At the same time, the data raise questions about how geographical differences arose and the direction of spread from one region to another. The answers to these questions will require further sampling of geographically diverse whole CMV genomes, CMV sequence data from ancient samples, or both. Finally, the data highlight that the geographical and allelic differences between proteins being trialled as potential vaccines needs to be considered when designing vaccines. Our findings raise the possibility that vaccines based on strain-specific gB or other viral antigens may fail to induce sufficient cross-protection globally against circulating variants.

Funding

OC is supported by a UKRI MRC grant “MR/N013867/1”.

CV and CA are funded by Wellcome Trust Grant No. “204870/Z/16/Z”.

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

SG is supported by operating grants from the Canadian Institutes of Health Research.

Conflicts of interest

SG receives consulting fees and research funding related to CMV from Moderna, Merck, GSK, VBI Vaccines, and Altona Diagnostics.

Code availability

The HMMcluster program is freely available under the MIT license, at https://github.com/ucl-pathgenomics/hmmcluster

Data availability

The accession identifiers (GenBank or SRA accessions) for the 259 genomes included in the main analysis and the genomes of those patients with self-reported ancestry are available in Dataset S1 and Dataset S2. The full alignment and full resolution phylogenetic trees are also available at https://github.com/ucl-pathgenomics/HCMV_resources_public.

Tables

Table 1. Recombination in multi-allelic regions.

Region	Genes	Number of breakpoints	NC_006273.2 coordinates of breakpoints (bp)
2	**RL5A; RL6**	**2**	**5907, 6290**
6	**RL11; RL12; RL13; UL1; UL2; UL4**	**2**	**12606, 13305**
8	**UL10; UL11; UL6; UL7; UL8; UL9**	**6**	**15751, 15860, 16390, 17553, 18419, 18672**
10	UL20; UL21A	2	26468, 26732
28	UL73; UL74	2	107640, 107848
42	UL116	1	166458
43	UL119; UL120; UL121	6	169045, 169283, 169475, 169525, 169782, 170046
50	UL144	2	182464, 182551
51	UL150A	1	183170
72	TRS1	4	231683, 231783, 232227, 232415

Each multi-allelic region was assessed for evidence of recombination. Firstly, each region was examined using a set of seven recombination methods implemented in RDP5. We then visually investigated the phylogenies either side of predicted breakpoints of those multi-allelic regions with evidence of recombination (a region was considered to have evidence of recombination if at least 5 methods in RDP5 were significant). We removed breakpoints that could be explained by sub-clade structure. We underline and highlight in bold those regions where most sequences in the multiple-sequence alignment showed the recombination breakpoints.

Materials and methods

Data retrieval

A python script using Biopython (80), specifically the Entrez module, was used to access the SRA and NCBI nucleotide databases for sequence information, and extract country and continent assignment for sequences.

Sequence assembly

SRA sequences for Zambian CMV genomes were downloaded using the SRA toolkit and assembled using an in house de novo assembly pipeline, which involves contig generation, optimal reference identification, scaffolding on to the reference sequence, and subsequent iterative mapping of NGS reads on the genome scaffold. These were then subject to haplotype reconstruction and relevant consensus sequences determined.

Ugandan sequences and historical clinical sequences of known ethnicity were also de novo assembled. Kenyan sequence data were assembled to a reference sequence using an in-house pipeline using the strain. Sequence positions with less than 10 read depth were labelled as n.

Haplotype reconstruction

Possible mixed infections were investigated with HaROLD, a tool for reconstructing haplotypes using co-varying variant frequencies in a probabilistic framework (31). HaROLD takes the bam files obtained from the assembly step (as explained above) and then reconstructs the optimal number of haplotypes for each sample. Haplotypes' sequences are then checked by reconstructing phylogenetic trees and are considered distinct if they have >2000 bp differences. We reconstructed a total of 17 haplotypes from 8 samples (1-3 haplotypes per sample). In line with the approach taken in de novo assembly, we ignored haplotypes with an average read depth of less than 10 bases (haplotype frequency * mean read depth).

Multiple sequence alignment

Multiple sequence alignments were obtained using MAFFT v7 (81), particularly variable sections were re-aligned using MUSCLE (82) and finished manually. Sequence alignments were viewed in the lightweight alignment viewer AliView (83). Alignments relative to a reference strain were only used to generate the heterozygosity per reference position calculation, these were generated using MAFFT with the " --add – keeplength" options, which allowed SNPs to be called based on differences to the reference Merlin (Refseq accession: NC_006273.2).

Measures of sequence diversity

Heterozygosity was generated using an in-house R function, using the following calculation. h is heterozygosity for a given polymorphic site with I alleles, such that the sum of all allele frequencies p equals 1. N is the number of sequences in the sample. Summing over all segregating sites S in an alignment, we get sum of site heterozygosity π.

$$h = \frac{n}{n-1}\,(1 - \sum_{j=1}^{S} p_i^2)\,\pi = \sum_{j=1}^{S} h_j$$

Multi-Dimensional Scaling

Pairwise distances were calculated using the dist.dna() with the nucleotide-nucleotide substitution matrix "TN93" (84) and with pairwise deletion by way of the R package Ape v.5.4 (85). Multi-dimensional scaling much like principal component analysis (PCA) is a method to attempt to simplify complex data into a more interpretable format, by reducing dimensionality of data whilst retaining most of the variation. In a genomics context we can use this on pairwise distance matrices, where each dimension is a sequence with data points of n-1 sequences pairwise distance. This allows us to observer patterns of population structure as “clusters”. MDS was implemented using the cmdscale() function with pairwise deletion in R (86).

Phylogenetic Reconstruction

Phylogenetic relationships of multi-allelic and example conserved regions in Figure 2 were constructed from nucleic acid sequences in IQ-TREE (87). Using a Maximum Likelihood GTR substitution model with a discrete Gamma heterogeneity model (88) and 1000 rounds of bootstrapping . We attempted to root the CMV homologous genome to the most suitable ex-CMV taxa Chimpanzee CMV (accession AF480884), however this outgroup was too far removed with distance >5 so unrooted trees were preferred. Trees were visualised using Figtree (89).

For the whole genome and concatenome phylogenetic analysis where CMV is known to recombine freely, a Neighbor-net split phylogenetic network analysis was undertaken using Splitstree version 4.1.5 (90). Non-default options chosen were HKY85 distance matrix, with equal site rate variation. Both terminal repeat regions were trimmed from alignments (although they had negligible impact) before analysis.

“F_{st}” F statistics

For calculating a F_{st} like statistic from sequence data, we can use the sum of site heterozygosity’s across a locus to produce γ_{st}. Where π_T is calculated as above using all samples in an alignment, π_S is an average of the same calculation for each sub population separately. (91).

$$\gamma_{ST} = \frac{\pi_T - \pi_S}{\pi_T}$$

Only sites with greater than 5% minor allele frequency were considered. To account for uneven African and European populations, either when defined by sampling location or when considering those sequences that are archetypally (90%> in admixture analyses) African or European, we repeatedly subsampled the European population to be equivalent to the number of African sequences (1000*) and took the mean of the site Fst’s.

When we multiply bootstrap sampled the 30 African and 30 European sequences, the mean number of pairwise differences for sequences within each population were determined, as well as the mean number of pairwise differences across all sequences. This can be used to estimate Fst in an efficient manner for multiple bootstraps (92).

Chi-squared analysis of allele proportions

For each region the allele assignments from HMMcluster were grouped by origin into African and European allele frequencies as observed in the admixture archetypal strains, which we tested for significant differences using a chi squared test of independence with the base R function chisq.test over allele frequencies. From the European data we generated expected allele frequencies which were compared against observed allele frequencies from Africa. A Benjamin-Hochberg adjusted False Discovery Rate (FDR) < 0.05 we determined as inferring significant distribution deviation where we assigned the multi-allelic region as “geographic”, otherwise they were labelled “pervasive”.

HMMcluster – Sequence clustering by Optimal Hidden Markov Models

We implemented a maximum likelihood allele assignment model in java on a Hidden Markov Models statistical framework. Briefly, this approach considers the genomic alignment as a set of contiguous blocks, within each block the model instantiates by perfectly representing each sequence as its own HMM, this results in the highest Likelihood but with an excessive number of parameters. Then the model considers the optimal way to combine HMM’s to keep the highest likelihood i.e. 259 models to 258, and continues to iterate with a greedy stepwise algorithm until only a single HMM is reached. A single HMM most poorly represents each sequence, such that the likelihood is lowest, but it uses the minimal number of parameters. To balance this likelihood and parameter problem, a typical approach is to appeal to the Akaike Information Criterion (AIC), and we use this here to identify the most parsimonious representation of a given genomic segment.

Consider that we have a set of sequences$\{x_{1j}, x_{2j} \dots x_{nj}\}$ where x_{ij} is the base found at position i in sequence j, where there are N positions in sequence ($1 \leq i \leq N$) and M sequences ($1 \leq j \leq M$). Each Hidden Markov Model, not considering insertion and deletion states (of which we ignore insertion states) is defined as a series of match states which are represented by the probability of the emissions from that state. That is the Hidden Markov Model is defined by where $p_i(x)$ is the probability that match state i emits base x, and $\sum_k p_j(x_k) = 1$.

$$\left\{ \begin{Bmatrix} p_1(x_1) \\ p_1(x_2) \\ \vdots \end{Bmatrix}, \begin{Bmatrix} p_2(x_1) \\ p_2(x_2) \\ \vdots \end{Bmatrix}, \begin{Bmatrix} p_3(x_1) \\ p_3(x_2) \\ \vdots \end{Bmatrix} \cdots \right\}$$

In this case, the probability that sequence j would arise from this hidden Markov model is equal to $\prod_i p_i(x_{ij})$ or the log likelihood is given by $\sum_i \log p_i(x_{ij})$. The total log likelihood for the set of M sequences is then equal to $\sum_j \sum_i \log p_i(x_{ij}) = \sum_i \sum_j \log p_i(x_{ij})$.

If we consider a given location i and imagine that at this site m_{i1} of the sequences have base x_1, m_{i2} of the sequences have base x_2, etc, with $\sum_k m_{ik} = M$, then we can sum over identities of bases rather than sum over sequences, and the log likelihood becomes $\sum_i \sum_k m_{ik} \log p_i(x_k)$. It turns out, not surprisingly, that the best (i.e. maximum likelihood) values for $p_i(x_k)$ is equal to the fraction of the sequences that have base x_k at that position, that is, $\hat{p}_i(x_k) = \frac{m_{ik}}{M}$. Substituting this in yields the highest likelihood of the set of sequences is given by $\sum_i \sum_k m_{ik} \log \frac{m_{ik}}{M}$.

We implemented the model to look for evidence of population structure initially within 200bp genome slices. Identified loci were refined by Maximum likelihood, and any overlapping regions concatenated, and again start / stop positions refined by maximum likelihood.

Recombination analysis

Genome sequences were examined for evidence of systematic recombination events using the Recombination Detection Program (RDP) version RDP5.5 with the maximum likelihood tree option (93). The RDP software includes a suite of recombination-detecting algorithms where we used seven, namely phylogenetic (RDP, BOOTSCAN and SISCAN) and substitution (GENECONV, MAXCHI, CHIMAERA, and 3-SEQ) methods to generate evidence of recombination. Using a Bonferroni corrected P-value cut-off of $\leqslant 0.05$ significant scores with 5 or more of the seven algorithms, found in a group of 4 or more non-haplotype sequences were considered significant and the phylogeny either side examined to determine if it was a true significant.

Population structure

Population structure was analysed in an unsupervised fashion with Admixture 1.3.0 (48). Alignments were converted to VCF format using snp-sites (94), and sites with minimum allele frequency < 5% were trimmed. Sequences were randomly subsampled to generate more proportionate sample sizes per continent (30 European, 30 African and both (2) Asian) 1000 time. For each of the 1000 sample draws admixture was run for a k ranging from 1 to 10 with 20-fold cross validation. As recommended in the admixture manual we thinned the markers according to the observed sample correlation coefficients using the plink argument “--indep-pairwise 50 10 0.1”. Analyses were visualised in R.

Identifying Epitopes

Known CMV B and T cell epitopes were downloaded from IEDB (51) then mapped to Merlin reference strain genomic co-ordinates by tblastn (95). Predicted epitopes were ignored. cmvdrg (96) was used to identify which variants are synonymous or nonsynonymous when translated. Sites with less than 10% variants in either African or European populations were ignored. As sites with low variability can still exhibit high Fst values, we limited the analysis to sites where the consensus base was different between the archetypal African and European sequences, this removed 11% of sites. This allowed variant sites to be analysed together from geographical, immune, and protein effect frames of reference.

References

1. M. Zuhair, *et al.*, Estimation of the worldwide seroprevalence of cytomegalovirus: A systematic review and meta-analysis. *Rev. Med. Virol.* **29**, e2034 (2019).

2. J. I. Cohen, Herpesvirus latency. *J. Clin. Invest.* **130**, 3361 (2020).

3. A. J. Davison, *et al.*, The Order Herpesvirales. *Arch. Virol.* **154**, 171–177 (2009).

4. S.-Y. Cho, D.-G. Lee, H.-J. Kim, Cytomegalovirus Infections after Hematopoietic Stem Cell Transplantation: Current Status and Future Immunotherapy. *Int. J. Mol. Sci.* **20** (2019).

5. C. Marsico, D. W. Kimberlin, Congenital Cytomegalovirus infection: advances and challenges in diagnosis, prevention and treatment. *Ital. J. Pediatr.* **43**, 38 (2017).

6. Institute of Medicine (US) Committee to Study Priorities for Vaccine Development, *Vaccines for the 21st Century: A Tool for Decisionmaking*, K. R. Stratton, J. S. Durch, R. S. Lawrence, Eds. (National Academies Press (US), 2000) (June 27, 2022).

7. A. Dolan, *et al.*, Genetic content of wild-type human cytomegalovirus. *J. Gen. Virol.* **85**, 1301–1312 (2004).

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

9. N. Renzette, B. Bhattacharjee, J. D. Jensen, L. Gibson, T. F. Kowalik, Extensive Genome-Wide Variability of Human Cytomegalovirus in Congenitally Infected Infants. *PLoS Pathog.* **7** (2011).

10. S. Sijmons, *et al.*, High-Throughput Analysis of Human Cytomegalovirus Genome Diversity Highlights the Widespread Occurrence of Gene-Disrupting Mutations and Pervasive Recombination. *J. Virol.* **89**, 7673–7695 (2015).

11. F. Lassalle, *et al.*, Islands of linkage in an ocean of pervasive recombination reveals two-speed evolution of human cytomegalovirus genomes. *Virus Evol.* **2** (2016).

12. S. Chou, Reactivation and Recombination of Multiple Cytomegalovirus Strains from Individual Organ Donors. *J. Infect. Dis.* **160**, 11–15 (1989).

13. E. Puchhammer-Stöckl, I. Görzer, Human cytomegalovirus: an enormous variety of strains and their possible clinical significance in the human host. *Future Virol.* **6**, 259–271 (2011).

14. E. Puchhammer-Stöckl, I. Görzer, Cytomegalovirus and Epstein-Barr virus subtypes—The search for clinical significance. *J. Clin. Virol.* **36**, 239–248 (2006).

15. U. Meyer-König, K. Ebert, B. Schrage, S. Pollak, F. T. Hufert, Simultaneous infection of healthy people with multiple human cytomegalovirus strains. *The Lancet* **352**, 1280–1281 (1998).

16. N. M. Suárez, *et al.*, Human Cytomegalovirus Genomes Sequenced Directly From Clinical Material: Variation, Multiple-Strain Infection, Recombination, and Gene Loss. *J. Infect. Dis.* **220**, 781–791 (2019).

17. N. M. Suárez, *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.* **220**, 792–801 (2019).

18. D. H. Shepp, *et al.*, Cytomegalovirus Glycoprotein B Groups Associated with Retinitis in AIDS. *J. Infect. Dis.* **174**, 184–187 (1996).

19. V. C. Emery, *et al.*, Differential decay kinetics of human cytomegalovirus glycoprotein B genotypes following antiviral chemotherapy. *J. Clin. Virol.* **54**, 56–60 (2012).

20. E. Paradowska, *et al.*, Distribution of cytomegalovirus gN variants and associated clinical sequelae in infants. *J. Clin. Virol.* **58**, 271–275 (2013).

21. S. Pignatelli, *et al.*, Cytomegalovirus gN Genotypes Distribution among Congenitally Infected Newborns and Their Relationship with Symptoms at Birth and Sequelae. *Clin. Infect. Dis.* **51**, 33–41 (2010).

22. E. Paradowska, *et al.*, Cytomegalovirus alpha-chemokine genotypes are associated with clinical manifestations in children with congenital or postnatal infections. *Virology* **462–463**, 207–217 (2014).

23. R. He, *et al.*, Sequence Variability of Human Cytomegalovirus UL146 and UL147 Genes in Low-Passage Clinical Isolates. *Intervirology* **49**, 215–223 (2006).

24. S. K. Pati, *et al.*, Genotypic Diversity and Mixed Infection in Newborn Disease and Hearing Loss in Congenital Cytomegalovirus Infection. *Pediatr. Infect. Dis. J.* **32**, 1050–1054 (2013).

25. A. Mozzi, *et al.*, Past and ongoing adaptation of human cytomegalovirus to its host. *PLOS Pathog.* **16**, e1008476 (2020).

26. A. J. Davison, *et al.*, The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome. *J. Gen. Virol.* **84**, 1053–1053 (2003).

27. S. Sijmons, *et al.*, A Method Enabling High-Throughput Sequencing of Human Cytomegalovirus Complete Genomes from Clinical Isolates. *PLOS ONE* **9**, e95501 (2014).

28. D. J. Dargan, *et al.*, Sequential mutations associated with adaptation of human cytomegalovirus to growth in cell culture. *J. Gen. Virol.* **91**, 1535–1546 (2010).

29. G. S. Jung, *et al.*, Full genome sequencing and analysis of human cytomegalovirus strain JHC isolated from a Korean patient. *Virus Res.* **156**, 113–120 (2011).

30. , Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).

31. C. Venturini, *et al.*, Haplotype assignment of longitudinal viral deep-sequencing data using co-variation of variant frequencies. *Virus Evol.*, veac093 (2022).

32. J. Pang, *et al.*, Mixed cytomegalovirus genotypes in HIV-positive mothers show compartmentalization and distinct patterns of transmission to infants. *eLife* **9**, e63199 (2020).

33. S. Gantt, *et al.*, Prospective Characterization of the Risk Factors for Transmission and Symptoms of Primary Human Herpesvirus Infections Among Ugandan Infants. *J. Infect. Dis.* **214**, 36–44 (2016).

34. D. P. Depledge, *et al.*, Deep Sequencing of Viral Genomes Provides Insight into the Evolution and Pathogenesis of Varicella Zoster Virus and Its Vaccine in Humans. *Mol. Biol. Evol.* **31**, 397–409 (2014).

35. N. M. Suárez, *et al.*, Whole-Genome Approach to Assessing Human Cytomegalovirus Dynamics in Transplant Patients Undergoing Antiviral Therapy. *Front. Cell. Infect. Microbiol.* **10** (2020).

36. C. Mattick, *et al.*, Linkage of human cytomegalovirus glycoprotein gO variant groups identified from worldwide clinical isolates with gN genotypes, implications for disease associations and evidence for N-terminal sites of positive selection. *Virology* **318**, 582–597 (2004).

37. P. Smyth, “Clustering Sequences with Hidden Markov Models” in *Advances in Neural Information Processing Systems 9*, M. C. Mozer, M. I. Jordan, T. Petsche, Eds. (MIT Press, 1997), pp. 648–654.

38. S. Correia, *et al.*, Sequence Variation of Epstein-Barr Virus: Viral Types, Geography, Codon Usage, and Diseases. *J. Virol.* **92**, e01132-18 (2018).

39. C. Grose, Pangaea and the Out-of-Africa Model of Varicella-Zoster Virus Evolution and Phylogeography. *J. Virol.* **86**, 9558–9565 (2012).

40. M. L. Szpara, *et al.*, Evolution and Diversity in Human Herpes Simplex Virus Genomes. *J. Virol.* **88**, 1209–1227 (2014).

41. M. Quinlivan, *et al.*, The Molecular Epidemiology of Varicella-Zoster Virus: Evidence for Geographic Segregation. *J. Infect. Dis.* **186**, 888–894 (2002).

42. J. Novembre, *et al.*, Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).

43. D. Gurdasani, *et al.*, The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).

44. S. F. Atabani, *et al.*, Cytomegalovirus Replication Kinetics in Solid Organ Transplant Recipients Managed by Preemptive Therapy. *Am. J. Transplant.* **12**, 2457–2464 (2012).

45. A. J. Bradley, *et al.*, Genotypic analysis of two hypervariable human cytomegalovirus genes. *J. Med. Virol.* **80**, 1615–1623 (2008).

46. S. Pignatelli, *et al.*, Human cytomegalovirus glycoprotein N (gpUL73-gN) genomic variants: identification of a novel subgroup, geographical distribution and evidence of positive selective pressure. *J. Gen. Virol.* **84**, 647–655 (2003).

47. S. Shringarpure, E. P. Xing, Effects of Sample Selection Bias on the Accuracy of Population Structure and Ancestry Inference. *G3 GenesGenomesGenetics* **4**, 901–911 (2014).

48. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

49. D. J. Lawson, L. van Dorp, D. Falush, A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat. Commun.* **9**, 3258 (2018).

50. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

51. R. Vita, *et al.*, The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).

52. E. Van Damme, M. Van Loock, Functional annotation of human cytomegalovirus gene products: an update. *Front. Microbiol.* **5** (2014).

53. G. W. G. Wilkinson, *et al.*, Human cytomegalovirus: taking the strain. *Med. Microbiol. Immunol. (Berl.)* **204**, 273–284 (2015).

54. M. R. Schleiss, Recombinant cytomegalovirus glycoprotein B vaccine: Rethinking the immunological basis of protection. *Proc. Natl. Acad. Sci.* **115**, 6110–6112 (2018).

55. I. Baraniak, *et al.*, Protection from cytomegalovirus viremia following glycoprotein B vaccination is not dependent on neutralizing antibodies. *Proc. Natl. Acad. Sci.* **115**, 6273–6278 (2018).

56. C. S. Nelson, *et al.*, HCMV glycoprotein B subunit vaccine efficacy mediated by nonneutralizing antibody effector functions. *Proc. Natl. Acad. Sci.* **115**, 6267–6272 (2018).

57. M. Foglierini, J. Marcandalli, L. Perez, HCMV Envelope Glycoprotein Diversity Demystified. *Front. Microbiol.* **10** (2019).

58. G. Guo, *et al.*, Polymorphisms and features of cytomegalovirus UL144 and UL146 in congenitally infected neonates with hepatic involvement. *PLoS ONE* **12**, e0171959 (2017).

59. C. Berg, *et al.*, The frequency of cytomegalovirus non-ELR UL146 genotypes in neonates with congenital CMV disease is comparable to strains in the background population. *BMC Infect. Dis.* **21**, 386 (2021).

60. J. J. C. de Vries, *et al.*, Rapid Genotyping of Cytomegalovirus in Dried Blood Spots by Multiplex Real-Time PCR Assays Targeting the Envelope Glycoprotein gB and gH Genes. *J. Clin. Microbiol.* **50**, 232–237 (2012).

61. H.-Y. Wang, *et al.*, Common Polymorphisms in the Glycoproteins of Human Cytomegalovirus and Associated Strain-Specific Immunity. *Viruses* **13**, 1106 (2021).

62. F. Zavaglio, *et al.*, Detection of Genotype-Specific Antibody Responses to Glycoproteins B and H in Primary and Non-Primary Human Cytomegalovirus Infections by Peptide-Based ELISA. *Viruses* **13**, 399 (2021).

63. W. Barrett-Muir, *et al.*, Genetic variation of varicella-zoster virus: evidence for geographical separation of strains. *J. Med. Virol.* **70 Suppl 1**, S42-47 (2003).

64. F. Wegner, F. Lassalle, D. P. Depledge, F. Balloux, J. Breuer, Coevolution of Sites under Immune Selection Shapes Epstein–Barr Virus Population Structure. *Mol. Biol. Evol.* **36**, 2512–2521 (2019).

65. C. Firth, *et al.*, Using Time-Structured Data to Estimate Evolutionary Rates of Double-Stranded DNA Viruses. *Mol. Biol. Evol.* **27**, 2038–2051 (2010).

66. M. Bates, A. B. Brantsaeter, Human cytomegalovirus (CMV) in Africa: a neglected but important pathogen. *J. Virus Erad.* **2**, 136–142 (2016).

67. U. A. Gompels, *et al.*, Human Cytomegalovirus Infant Infection Adversely Affects Growth and Development in Maternally HIV-Exposed and Unexposed Infants in Zambia. *Clin. Infect. Dis.* **54**, 434–442 (2012).

68. R. F. Pass, B. Anderson, Mother-to-Child Transmission of Cytomegalovirus and Prevention of Congenital Infection. *J. Pediatr. Infect. Dis. Soc.* **3**, S2–S6 (2014).

69. F. Ribeiro da Silva, The slave trade and the development of the Atlantic Africa port system, 1400s–1800s. *Int. J. Marit. Hist.* **29**, 138–154 (2017).

70. N. Renzette, *et al.*, Limits and patterns of cytomegalovirus genomic diversity in humans. *Proc. Natl. Acad. Sci.* **112**, E4120–E4128 (2015).

71. M. Guellil, *et al.*, "Ancient herpes simplex 1 genomes reveal recent viral structure in Eurasia" (Genomics, 2022) https:/doi.org/10.1101/2022.01.19.476912 (January 27, 2022).

72. J. B. Plotkin, J. Dushoff, Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proc. Natl. Acad. Sci.* **100**, 7152–7157 (2003).

73. S. F. Elena, R. Miralles, A. Moya, Frequency-Dependent Selection in a Mammalian RNA Virus. *Evolution* **51**, 984–987 (1997).

74. M. Zhou, J.-M. Lanchy, B. J. Ryckman, Human Cytomegalovirus gH/gL/gO Promotes the Fusion Step of Entry into All Cell Types, whereas gH/gL/UL128-131 Broadens Virus Tropism through a Distinct Mechanism. *J. Virol.* **89**, 8999–9009 (2015).

75. J. Kalser, *et al.*, Differences in Growth Properties among Two Human Cytomegalovirus Glycoprotein O Genotypes. *Front. Microbiol.* **8**, 1609 (2017).

76. S. L. Fishman, A. D. Branch, The Quasispecies Nature and Biological Implications of the Hepatitis C Virus. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **9**, 1158–1167 (2009).

77. A. M. Di Bisceglie, *et al.*, AGE, RACE AND VIRAL GENOTYPE ARE ASSOCIATED WITH THE PREVALENCE OF HEPATITIS B E ANTIGEN IN CHILDREN AND ADULTS WITH CHRONIC HEPATITIS B. *J. Viral Hepat.* **26**, 856–865 (2019).

78. L. Ségurel, Z. Gao, M. Przeworski, Ancestry runs deeper than blood: The evolutionary history of ABO points to cryptic variation of functional importance. *Bioessays* **35**, 862–867 (2013).

79. S. John, *et al.*, Multi-antigenic human cytomegalovirus mRNA vaccines that elicit potent humoral and cell-mediated immunity. *Vaccine* **36**, 1689–1699 (2018).

80. P. J. A. Cock, *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

81. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

82. R. C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

83. A. Larsson, AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).

84. K. Tamura, M. Nei, Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526 (1993).

85. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

86. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2014).

87. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

88. Z. Yang, Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).

89. A. Rambaut, FigTree v1.4.2, A Graphical Viewer of Phylogenetic Trees (2018).

90. D. H. Huson, SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).

91. R. Chakraborty, M. Nei, Genetic differentiation of quantitative characters between populations or species: I. Mutation and random genetic drift. *Genet. Res.* **39**, 303–314 (1982).

92. R. R. Hudson, M. Slatkin, W. P. Maddison, Estimation of Levels of Gene Flow from DNA Sequence Data. *Genetics* **132**, 583–589 (1992).

93. D. P. Martin, B. Murrell, M. Golden, A. Khoosal, B. Muhire, RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1** (2015).

94. A. J. Page, *et al.*, SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genomics* **2** (2016).

95. C. Camacho, *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

96. O. J. Charles, C. Venturini, J. Breuer, "cmvdrg - An R package for Human Cytomegalovirus antiviral Drug Resistance Genotyping" (Bioinformatics, 2020) https:/doi.org/10.1101/2020.05.15.097907 (November 13, 2020).

Figure Legends

Figure 1. Circular genome map showing nucleotide diversity and multi-allelic regions. Tracks numbered from in to out: (Track1 - red) Barplot of nucleotide diversity (calculated as heterozygosity) is shown as bars of Heterozygosity (red) along a 50bp moving average; (Track2) Open Reading Frames in the CMV genome are coloured by gene family as defined by bottom legend and in (26); (Track3 – blue) Multi-allelic regions as defined using HMMcluster are highlighted in translucent blue; (Track4) ORF names. We also show a representative multi-sequence alignment for conserved (left) and multi-allelic (right) regions.

Figure 2. Unrooted Maximum Likelihood phylogenetic trees of representative multi-allelic and conserved regions. Tips were grouped if within 5% of the maximum taxa distance and are shown as triangles where size indicates the number of grouped sequences and colour represents different allele from HMMcluster. Small hard to see fans have been blown up and are represented by fans within circles. Nodes with bootstrap support >90% are shown as red diamonds. **Note:** Scale bars differ for each figure. **A)** Multi-allelic region 2 (RL5A RL6) (5 alleles). **B)** Multi-allelic region 30 (UL75) (2 alleles). **C)** Example conserved region (UL105) (1 allele) of comparable alignment length. Variability of C is much less than A and B with no support for HMM derived clusters**. D)** A and C when drawn to the scale of B, the example conserved region tree becomes difficult to see at this representation reflecting the relatively minor variation it encodes. Sequences with greater than 15% ambiguous bases were removed before phylogenetic reconstruction.

Figure 3 Multi-Dimensional Scaling of all CMV genomes. The figure shows multidimensional scale analysis for all CMV strains analysed (n=259): each dot represents a CMV strain, and the colour indicates the continent of isolation (Europe includes European and Middle eastern genomes). The analysis was done in three scenarios: A) whole genome, B) conserved regions (conserved concatenome), C) multi-allelic regions (multi-allelic concatenome). This analysis shows an overall trend for geographical segregation for the whole genome (A) and the conserved regions (C), but not for the multi-allelic regions (B).

Figure 4. Admixture analysis in CMV's conserved regions. Admixture inferred ancestral lineages reconstructed from CMV conserved concatenomes support evidence of geographic segregation in CMV. The admixtures derived from a representative K=2 model of 62 sequences, were projected to the remaining 197 sequences. A) This plot shows admixture proportions for whole CMV dataset

(n=259 strains) grouped by continent. B) The red and blue cluster components were used to colour sequences in the conserved concatenome MDS. Select common reference strains have been labelled.