Evolution of the *Cytomegalovirus* RL11 Gene Family in Old World monkeys and Great Apes

Ulad Litvin[1*], Eddie C.Y. Wang[2], Richard J. Stanton[2], Ceri A. Fielding[2], Joseph Hughes[1]

1 – MRC-University of Glasgow Centre for Virus Research, Glasgow G61 1QH, UK

2 – Infection and Immunity, Cardiff University School of Medicine, Cardiff CF14 4XN, UK

*Corresponding author: E-mail: u.litvin.1@research.gla.ac.uk

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

Abstract

Cytomegalovirus is a genus of herpesviruses, members of which share a long history of co-evolution with their primate hosts including Great Apes, Old and New World monkeys. These viruses are ubiquitous within their host populations and establish lifelong infection in most individuals. Although asymptomatic in healthy individuals, infection poses a significant risk to individuals with a weakened or underdeveloped immune system. The genome of human cytomegalovirus is the largest among human-infecting viruses, and comprises at least fifteen separate gene families, which may have arisen by gene duplication. Within human cytomegalovirus, the RL11 gene family is the largest. RL11 genes are non-essential *in vitro* but have immune evasion roles that are likely critical to persistence *in vivo.* These genes

The Author(s) 2024. Published by Oxford University Press.

demonstrate an extreme level of inter-species and intra-strain sequence diversity, that makes it challenging to deduce the evolutionary relationships within this gene family. Understanding the evolutionary relationships of these genes, especially accurate ortholog identification, is essential for reconstructing ancestral genomes, deciphering gene repertoire and order, and enabling reliable functional analyses across the *Cytomegalovirus* species, thereby offering insights into evolutionary processes, genetic diversity, and the functional significance of genes. In this work, we combined *in silico* genome screening with sequence-based and structure-guided phylogenetic analysis to reconstruct the evolutionary history of the RL11 gene family. We confirmed that RL11 genes are unique to cytomegaloviruses of Old World monkeys and Great Apes, showing that this gene family was formed by multiple early duplication events and later lineage-specific losses. We identified four main clades of RL11 genes and showed that their expansions were mainly lineage-specific and happened independently in cytomegaloviruses of Great Apes, African Old World monkeys and Asian Old World monkeys. We also identified groups of orthologous genes across the *Cytomegalovirus* tree showing that some human cytomegalovirus-specific RL11 genes emerged before the divergence of human and chimpanzee cytomegaloviruses but were subsequently lost in the latter. The extensive and dynamic species-specific evolution of this gene family suggests their functions target elements of host immunity that have similarly co-evolved during speciation.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

Introduction

The cytomegaloviruses (CMVs, genus *Cytomegalovirus*, subfamily *Betaherpesvirinae*, family *Orthoherpesviridae*, order *Herpesvirales*, Lefkowitz et al. 2018) are a group of herpesviruses that infect only primate hosts. Based on the similar topologies of *Cytomegalovirus* and primate phylogenetic trees and limited cross-species transmissions, CMVs are thought to have co-speciated with their hosts and now are restricted to them (Brito et al. 2021; McGeoch et al. 2006). However, some CMVs did switch hosts in the past several million years (Brito et al. 2021; Murthy et al. 2019). Currently, at least eleven primate species are known to be infected with species-specific CMVs (Cagliani et al. 2020), of which human CMV (HCMV, *Human betaherpesvirus 5*) is the best studied. Like other members of the *Orthoherpesviridae* family, primary infection with HCMV leads to the establishment of a lifelong latent infection, despite the induction of robust humoral and cellular immunity. Prior infection and the presence of latent virus is insufficient to prevent superinfection, thus individuals can carry multiple strains simultaneously. Primary infection, reactivation, and reinfection, are usually asymptomatic in healthy individuals (de la Hoz et al. 2002). However, HCMV poses a significant risk to people with compromised (transplant recipients undergoing immunosuppressive treatment, individuals with HIV co-infection) or immature (infants) immune systems leading to high viral loads and end-organ disease (Griffiths & Reeves 2021). According to seroprevalence studies between 78% and 88% of the adult population worldwide is infected with HCMV, however the percentage varies with geographical region and age group (Zuhair et al. 2019).

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

HCMV is an enveloped DNA virus. It has a linear double stranded DNA genome of around 235 kilobase pairs, one of the largest among viruses infecting animals. The genome is composed of two regions flanked by pairs of inverted repeats: a long unique region (U_L) and a short unique region (U_S). It carries approximately 170 canonical protein coding genes, most of which can be grouped into fifteen gene families (Davison, Dolan, et al. 2003). The RL11 gene family is the largest and the most diverse gene family in the HCMV genome. It encompasses fourteen genes (RL5A, RL6, RL11 – RL13, UL1, UL4 – UL11) located at the end of the UL next to the left terminal repeat. RL11 genes encode type 1 membrane glycoproteins with comparable domain architecture (Davison, Akter, et al. 2003). Most of these proteins carry a signal peptide (SP), an extracellular region similar to immunoglobulin variable domain (IgV-like domain) known as RL11D, and a transmembrane domain (TMD). However, certain regions can be missing from some proteins, e.g. RL5A and RL6 genes encode proteins without a SP and TMD, UL4 lacks a TMD and is secreted (Vlachava et al. 2023), while UL5 encodes only a TMD. Orthologous genes belonging to the RL11 family show the highest level of divergence between different HCMV genotypes of any genes in the genome (Sekulin et al. 2007). It has been reported that the RL11 gene family might have originated as the result of the primate CD229/SLAMF3/LY9 gene being co-opted into the CMV genome (Engel et al. 2011).

All members of the RL11 gene family are dispensable for HCMV reproduction *in vitro* (Atalay et al. 2002), however the glycoproteins encoded by these genes play critical roles in the regulation of the host immune response *in vivo*. The RL11, RL12, and RL13 genes encode membrane glycoproteins which bind the Fc region of IgG and prevent IgG-mediated clearance of the virus (Cortese et al. 2012; Lilley et al. 2001). Proteins UL7 and UL8 interact

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

with a receptor on the surface of dendritic cells and activated neutrophils and impair their ability to secrete proinflammatory cytokines (Engel et al. 2011; Pérez-Carmona et al. 2018). Proteins UL10 and UL11 suppress proliferation and proinflammatory cytokine production in T cells (Bruno et al. 2016; Gabaev et al. 2011), while UL4 binds to TRAIL to prevent NK-cell activation and TRAIL-mediated apoptosis (Vlachava et al. 2023). Thus, RL11 proteins perform a wide range of functions that are likely critical to evading and manipulating host immunity and promoting lifelong persistence of the virus.

To date, complete CMV genomes from ten different species have been assembled. From these studies we know that chimpanzee CMV, *Panine betaherpesvirus 2*, has eleven RL11 genes (Davison, Dolan, et al. 2003); African green monkey CMV, *Cercopithecine betaherpesvirus 5*, has nineteen (Davison et al. 2013); rhesus macaque CMV, *Macacine betaherpesvirus 3*, and Japanese macaque CMV have twenty RL11 genes each (Davison et al. 2013; Taher et al. 2020). In contrast, CMVs of New-World Monkeys (NWM) – owl monkey CMV, *Aotine betaherpesvirus 1*, and squirrel monkey CMV, *Saimiriine betaherpesvirus 4*, seem to lack RL11 genes (Davison et al. 2013). The same is true for the closely related betaherpesviruses of rodents (mouse, rat and tupaia); their genomes do not carry RL11 genes (Davison, Akter, et al. 2003). Genomes of several other Old World monkeys (OWM) CMVs (*Papiine betaherpesvirus 4*, *Mandrilline betaherpesvirus 5* and *Macacine betaherpesvirus 8*) are assembled (Blewett et al. 2015; Marsh et al. 2011) but the number of RL11 genes in these genomes is not clear because of the poor annotation of RL11 protein-coding genes, at least in part due to the extreme level of sequence divergence seen amongst this gene family.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

It was reported that CR1 genes of human adenoviruses, encoded by the E3 genomic region, are potentially related to the RL11 gene family. CR1 genes also encode type 1 membrane glycoproteins with a SP, up to three Ig-like domains called CR1 domains and a TMD (Davison, Akter, et al. 2003). These genes, similar to members of the RL11 gene family, are dispensable for virus reproduction *in vitro* (Hitt et al. 1995), but not much is known about their role *in vivo* apart from the fact that they assist other genes located in the E3 region to perform immune evasion functions (Singh et al. 2013). Multiple sequence alignments show that CR1 and RL11D domain regions form around fifteen relatively conserved residues including one tryptophan and two cysteines present in the majority of proteins encoded by these genes (Davison, Akter, et al. 2003). CR1 genes also show significant diversity between different genotypes and can be truncated and/or lack one or more of their functional regions (Jacobs et al. 2004). However, the phylogenetic relationships between these genes and the RL11 gene family are currently unclear.

To our knowledge the evolution of the RL11 gene family has not yet been investigated. Because of the low sequence conservation between members of this gene family and relatively short length of the encoded proteins, elucidation of the phylogenetic relationships between these genes presents a challenging task. With recent advances in the field of protein structure prediction and availability of new tools which help with the inference of structure-aware phylogenies, many evolutionary questions that could not be answered before with pure sequence-based approaches can be revisited. Due to the poor annotation of some genomes and to collate a comprehensive dataset of RL11 genes, we performed a systematic *in silico* screening of genomes belonging to CMVs, related betaherpesviruses, mastadenoviruses and their mammalian hosts. We conducted a traditional sequence-based

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

and structure-guided phylogenetic analysis for the discovered members of the RL11 gene family to identify robust orthologs and guide our understanding of functional information within the gene family.

Materials and methods

In silico genome screening

A set of thirty-four betaherpesvirus genomes, twenty one *Mastadenovirus* genomes, and twenty mammalian host genomes (see Supplementary Table 1) were analysed using database-integrated genome screening (DIGS) software v.2.0 (Blanco-Melo et al. 2023). DIGS is a genome screening pipeline based on Basic Local Alignment Search Tool (BLAST) that allows users to perform iterative searches over genomes of interest using either protein or nucleotide probes and stores hits in a relational database. We conducted three screening iterations (see Figure S1) on the same set of genomes, using empirically chosen thresholds based on multiple DIGS test runs. A tblastn bitscore threshold of 31 ensured a high diversity of nucleotide hits while minimizing false positives. A sequence length threshold of 85 bp captured core regions of the IgV-like domain or TMD, excluding shorter hits. A defragment range of 20 bp allowed the merging of two closely located hits into one while preventing the fusion of two nearby genes into a single hit. For the first search iteration we used a probe set of twenty annotated RL11 protein sequences from *Macacine betaherpesvirus 3* strain 68-1 BAC (JQ795930.1) and fourteen RL11 protein sequences from *Human betaherpesvirus 5* strain Merlin (AY446894.2). Although, *Human betaherpesvirus 5* UL8 protein is known to be the result of splicing between the UL7 gene and a downstream located ORF that encodes an

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

alternative TMD, in this study we used separate probes for the UL7 gene and the downstream ORF with alternative TMD which we refer to as UL8 in the text. Coding sequences of newly discovered RL11 genes were translated *in silico* and added to the probe set used for subsequent search iterations. The RL11 proteins with tblastn bitscore above 45 and sequence length above 75 aa (see Supplementary Table 2) discovered after the third iteration of screening were subjected to protein domain annotation, protein structure prediction and phylogenetic analysis.

Prediction of protein domains, functional regions, and glycosylation sites

We performed annotation of functional regions and protein domains for all *in silico* identified proteins using the web version of InterProScan 5 (Jones et al. 2014) in January 2024. Transmembrane domains were annotated based on TMHMM TMhelix coordinates from the InterProScan 5 results. For more accurate prediction of signal peptides, we submitted the same set of protein sequences to SignalP 6.0 (Teufel et al. 2022) in January 2024 with slow model mode and specifying Eukarya as an organism of choice. IgV-like domains could not be consistently predicted using InterProScan 5, therefore for their annotation we relied on multiple sequence alignments and the predicted protein structures. We defined the IgV-like domain as a region encompassing cysteine 398 and valine 588 (MAFFT alignment of RL11 proteins) which consists of nine beta strands and forms an IgV-like fold. To predict N-linked and O-linked glycosylation sites, we submitted the RL11 protein set to NetNGlyc 1.0 (Gupta & Brunak 2002) and NetOGlyc 4.0 (Steentoft et al. 2013) in January 2024.

Protein structure prediction and structure-guided alignment

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

Protein structures for all 160 *in silico* identified proteins were predicted using local ColabFold v.1.5.3 (Mirdita et al. 2022) and ESMFold v.1.0.3 (Lin et al. 2023) software with default settings. ColabFold generally outperforms ESMFold in predicting high-quality protein structures. However, its performance depends on multiple sequence alignments, leading to lower accuracy when publicly available databases lack sequences similar to the target protein. In contrast, ESMFold is a single-sequence structure predictor. It utilises an ESM-2 transformer protein language model that does not require multiple sequence alignments. This approach is advantageous for predicting proteins with low sequence similarity to publicly available sequences. In this study, we used a cutoff of an average pLDDT (predicted local distance difference test) score of 50, discarding structures of very low quality. Based on this criterion, 150 ESMFold and 130 ColabFold structures were considered suitable for the analysis. To maintain the diversity of protein structures in our analysis, we focused on the larger set of ESMFold predictions. Protein structures for representative members of each gene family and RL11 family clade with the highest pLDDT score were submitted to the PDBsum web server (Laskowski et al. 2018) in February 2024 to produce the topology maps of IgV-like domains.

Structure-guided pairwise sequence alignments of RL11 proteins were generated by Foldseek v.8.ef4e960 (van Kempen et al. 2023) using easy-search with 3Di+AA option. It allowed us to produce pairwise local sequence alignments of protein regions based on their 3D structure for each pair of proteins. The structure with the highest average LDDT score across all pairwise structure alignments (*Mandrilline betaherpesvirus 1* RL11J) was used as a reference to create a joint structure-guided multiple sequence alignment.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

Phylogenetic analysis

Multiple sequence alignments were obtained for all RL11 proteins using MAFFT v.7.475 (Katoh & Standley 2013) with default auto settings. To preserve only reliable homologous regions, MAFFT and Foldseek alignments were processed with ClipKIT v.1.3.0 (Steenwyk et al. 2020) using default smart-gap option. The phylogenies were inferred from the ClipKIT processed sequence alignments using IQ-TREE v.2.1.3 (Minh et al. 2020) with 100 transfer bootstrap replicates (Lemoine et al. 2018).

Phylogenetic tree reconciliation

To determine when RL11 gene duplication events and losses took place during the evolution of CMVs, we used GeneRax v.2.0.4 (Morel et al. 2020) to perform reconciliation of the RL11 gene family phylogeny generated using the ClipKIT processed MAFFT alignment and *Cytomegalovirus* phylogeny obtained from the ICTV web site (Lefkowitz et al. 2018). GeneRax allows users to produce species tree-aware phylogenies for gene families taking into account duplications, losses and horizontal gene transfer events. We used GeneRax with rec-model UndatedDL option.

Results

Systematic *in silico* genome screening

We performed a systematic *in silico* screening of betaherpesvirus, *Mastadenovirus* and mammalian genomes using a set of thirty-four annotated RL11 protein sequences from *Human betaherpesvirus 5* and *Macacine betaherpesvirus 3* as probes. The screening

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

revealed 141 RL11 genes in the genomes of OWM CMVs and Great Ape (GA) CMVs, seventeen CR1 genes (CR1-β, CR1-γ and CR1-δ) in the *Mastadenovirus* genomes and three genes from the EE50 gene family in the genomes of *Elephantid betaherpesvirus 1* and *Elephantid betaherpesvirus 5* (see Figure S2). The highest number of RL11 genes (twenty genes) was found in the *Papiine betaherpesvirus 4* genome, while the lowest number (ten genes) came from the *Panine betaherpesvirus 2* genome. The *Panine betaherpesvirus 2* UL4 gene was not found in our screen. Also, we did not find any RL11-related genes in the genomes of the mammalian hosts. All the genes, apart from UL5, UL8 (ORF downstream of UL7, see Materials and methods) and RL11N, encoded proteins with at least one extracellular IgV-like domain. Some CR1 genes encoded proteins with two or even three consecutive Ig-like domains in their extracellular region (Figure S3).

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

Phylogeny of the RL11, CR1 and EE50 gene families

We performed a phylogenetic analysis to clarify the evolutionary relationships between the RL11, CR1 and EE50 gene families (see Figure S4). According to the phylogeny, CR1 genes and EE50 genes formed a clade with UL7 and RL11P genes from the RL11 family, while a single CR1-δ gene from *Human adenovirus 4* was found within the RL11-γ clade of RL11 genes. However, because of the low bootstrap support of underlying nodes, the meaning of these phylogenetic relationships remains uncertain. To resolve this ambiguity, we predicted protein structures for all RL11, CR1 and EE50 proteins and produced topology maps for IgV-like domains from proteins of interest (Figure 1). The topology diagrams show that IgV-like domains of analysed proteins usually consist of two β-sheets with six β-strands in one sheet and three β-strands in the other with a conserved tryptophan residue in the β-strand C. Occasionally the region corresponding to the sixth β-strand C'' can be absent completely

(*Human adenovirus 4* CR1-δ) or be present but not form a β-strand (*Human adenovirus 4* CR1-γ).

All RL11 proteins together with some CR1-β and CR1-γ proteins also share a conserved disulfide bond between the β-strand C’ and β-strand D that connects two β-sheets together. However, in contrast to RL11, CR1 proteins usually have more than one Ig-like domain. CR1-δ proteins have a disulfide bond in a similar place but between the β-strand D and a short α-helix downstream of the β-strands C’. CR1 domains of *Human adenovirus 4* CR1-δ and *Human adenovirus 3* CR1-δ proteins adopt analogous folds, so it is not clear why one protein was grouped together with other CR1-β and CR1-γ proteins while the other was placed together with RL11 proteins (Figure S4). It is also surprising that UL7 and RL11P proteins form a clade with EE50 proteins. EE50 proteins of elephantid betaherpesviruses lack cysteine residues in the conserved positions of the IgV-like domain, a distinct mark of proteins encoded by CMV RL11 genes and *Mastadenovirus* CR1 genes. Therefore, the IgV-like domains of EE50 proteins lack disulfide bonds while proteins UL7/RL11P share a conserved disulfide bond between the β-strand C’ and β-strand D with other RL11 proteins (Figure 1). These relationships are highly unstable in different phylogenetic reconstruction methods (sequence versus structure-aware, see Figure S5), and suggests that branches with significantly higher substitution rates are being pulled together in an artefact called long branch attraction.

Phylogeny of the RL11 gene family

Since even using protein structure information we were unable to confidently establish the phylogenetic relationships between RL11, CR1 and EE50 gene families, we decided to focus

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

our phylogenetic and synteny analyses on the RL11 genes of CMVs. We performed a sequence-based and structure-guided phylogenetic analysis on the RL11 proteins (see Figure S5). Both phylogenies are broadly in agreement with the RL11, CR1, EE50 combined phylogeny (Figure S4) and show that the RL11 gene family diversified into four major clades which we called RL11-α, RL11-β, RL11-γ and RL11-δ. However, due to the lack of high bootstrap support basally in the phylogeny, the relationships between these major clades are difficult to ascertain with confidence. Three of these clades (RL11-α, RL11-γ and RL11-δ) have genes from both OWM CMVs and GA CMVs, while one clade (RL11-β) is unique to OWM CMVs. Both phylogenies consistently share the same sets of genes across the corresponding clades, apart from the unstable placement of the UL7 and RL11P genes which belong to the RL11-γ clade in accordance with the structure-guided phylogeny or to the RL11-δ clade according to the sequence-based phylogeny.

According to the consensus between the two phylogenies, the RL11-α clade encompasses six genes from the GA CMVs (RL5A, RL6, RL11, RL12, RL13 and UL5) and three genes (RL11A, RL11G and RL11T) from the OWM CMVs. RL11-β clade is unique to the OWM CMVs and includes five genes: RL11B, RL11C, RL11D, RL11E and RL11F. RL11-γ clade is composed of two genes (UL4 and UL6) from GA CMVs and four genes (RL11H, RL11I, RL11J and RL11R) from the OWM CMVs. RL11-δ clade is the largest clade among the four, and it consists of five genes (UL1, UL8, UL9, UL10 and UL11) from the GA CMVs and eight genes (RL11K, RL11L, RL11M, RL11N, RL11O, RL11O2, RL11Q and RL11S) from the OWM CMVs.

Domain organisation of RL11 proteins

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

In addition to the phylogenetic analysis, we performed an *in silico* prediction of functional regions for proteins encoded by the CMV RL11 genes. Although most of the genes encode proteins with a SP, IgV-like domain (RL11D) and a TMD, many genes lack either one or two of these functional regions (Figure 2). Since genes without some of these regions are related not to each other but to genes where all domains are present, we concluded that the loss of the functional regions occurred independently in different gene lineages. For instance, genes UL8 and RL11N encode only a TMD, but these genes are paralogous to UL9 and RL11O/RL11Q genes, respectively, encoding proteins with all three functional regions (SP, RL11D and TMD). Another example involves genes RL5A and RL6 which encode only the RL11D region. These *Human betaherpesvirus 5*-specific genes are coorthologs (genes that emerged as a result of a lineage-specific duplication event and share orthology to one or more genes from another lineage) of the RL11A gene which can be found in the OWM CMVs and encodes a SP, RL11D and TMD. The same is true for the *Human betaherpesvirus 5*-specific gene UL4 which does not encode a TMD, but its coorthologs RL11H and RL11I do, and gene UL6 which lacks a SP, but its ortholog RL11J has a SP region.

Since RL11 genes encode transmembrane proteins, extracellular regions of these proteins are likely to be glycosylated. We conducted an *in silico* prediction of glycosylation patterns for all members of the RL11 protein family and noticed that N-linked and O-linked glycosylations are usually found in distinct regions of the protein. In general, N-linked glycosylation sites are enriched in the RL11D regions, while O-linked glycosylation sites are concentrated upstream and/or downstream of RL11D. In the context of glycosylation, three genes are of particular interest: RL12 and its ortholog RL11T have a long heavily O- and N-

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

glycosylated region upstream of the RL11D, while RL11S specific to the OWM CMVs has a long heavily O-glycosylated region downstream of the RL11D.

Duplications and losses in the RL11 gene family

Analysis of the syntenic regions of the CMV genomes (Figure 3) shows that although NWM CMVs (*Saimiriine betaherpesvirus 4*, *Aotine betaherpesvirus 1*) do not possess any members of the RL11 gene family in this region, they share genes flanking RL11 members (RL1 and UL14) with other CMVs. We also performed reconciliation of the RL11 gene phylogeny with the species phylogeny of CMVs (see Figure S6) and added numbers of duplications and losses obtained from this analysis to the synteny diagram. The reconciliation analysis demonstrates that the RL11 gene family was shaped by several duplication events early in the evolution of OWM CMVs and GA CMVs, while losses of RL11 genes are generally more recent and usually lineage specific. For instance, ten duplication events occurred before the divergence of OWM CMVs and GA CMVs and led to the formation of the four major clades of RL11 genes (RL11-α, RL11-β, RL11-γ and RL11-δ).

Subsequent duplication events led to the expansion of different RL11 clades in different lineages of CMVs. The RL11-α clade expanded most significantly in the GA CMVs: *Human betaherpesvirus 5* genome contains six RL11-α genes while genomes of other CMVs generally have three RL11-α genes. RL11-β clade duplicated in the lineage of African OWM CMVs: the genome of *Papiine betaherpesvirus 4* carries nine RL11-β genes while other CMVs usually have no more than five. RL11-γ and RL11-δ clades have the highest number of members in the Asian OWM CMVs: these CMVs acquired the RL11I gene through the duplication of the RL11H, the RL11L gene – via the RL11K duplication, and the RL11O2 gene

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

after the duplication of either RL11O or RL11Q. However, genes related to the RL11M (RL11-δ clade) also expanded significantly in the GA CMVs leading to the emergence of UL8, UL9 and UL11 genes in *Human betaherpesvirus 5* and *Panine betaherpesvirus 2*. Interestingly, *Macacine betaherpesvirus 8* does not have a genomic region encompassing three RL11-α and -β genes (RL11E, RL11F and RL11G) which are present in the *Macacine betaherpesvirus 3* and Japanese macaque CMV. The RL11O2 gene is unique to the *Macacine betaherpesvirus 3* and *Macacine betaherpesvirus 8*; it is found at the position of the RL11P gene, which is missing in these CMVs.

Multiple rounds of lineage-specific duplication events make the identification of orthologs between OWM CMVs and GA CMVs particularly challenging. Based on the branches with high bootstrap support (Figure 2), we concluded that genes UL6 and RL11J have one-to-one orthologous relationship in all CMV genomes. This is similar to genes UL4 and RL11H, although in some CMVs the duplication of RL11H gene led to the emergence of the RL11I gene, coortholog of UL4. Genes RL12 and RL13 found in the *Human betaherpesvirus 5* and *Panine betaherpesvirus 2* are likely coorthologs of the RL11T gene specific to the OWM CMVs, while *Human betaherpesvirus 5*-specific genes RL5A and RL6 are coorthologs of the RL11A also found in the OWM CMVs. Orthologous relationships between other genes are less evident. UL7 may be an ortholog of the RL11P, genes UL8, UL9 and UL11 are likely related to the RL11M, and genes UL1 and UL10 have a close relationship to genes RL11O, RL11O2, RL11Q.

Another interesting observation is the distinct location of the RL11T gene in the genomes of the OWM CMVs. Unlike the rest of the RL11 genes, which are located in a row next to the

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

left terminal repeat of the U_L region, RL11T is situated next to the right end of the U_L region close to the U_S region. The orientation of the RL11T also does not match the orientation of the other RL11 genes. *Human betaherpesvirus 5*-specific genes RL5A and RL6 although located at the same place as their ortholog RL11A in the OWM CMVs, have opposite orientations.

Discussion

In this study, we performed a systematic *in silico* genome screening of CMVs, related betaherpesviruses, mastadenoviruses and their mammalian hosts. We found RL11 genes in all analysed OWM and GA CMV genomes. The numbers of identified RL11 genes match previously reported data for *Human betaherpesvirus 5*, *Cercopithecine betaherpesvirus 5*, *Macacine betaherpesvirus 3* and Japanese macaque CMV (Davison et al. 2013; Taher et al. 2020). The gene annotated as UL4 in the *Panine betaherpesvirus 2* genome was not detected in our analysis, indicating that this gene is a positional orthologue which is either unrelated to the RL11 family or pseudogenised beyond recognition. We also showed for the first time that *Papiine betaherpesvirus 4*, *Mandrilline betaherpesvirus 5* and *Macacine betaherpesvirus 8* have twenty-two, nineteen and seventeen RL11 genes respectively. We did not find RL11 genes in NWM CMV genomes nor in genomes of closely related betaherpesviruses (genera *Quwivirus*, *Muromegalovirus*, *Roseolovirus*) confirming that these viruses are lacking RL11 genes (Davison, Akter, et al. 2003).

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

It is worth mentioning that our genome screening approach was designed to find genes that encode proteins with noticeable sequence similarity to RL11 proteins. It was not designed to find all genes that encode proteins with structural similarity to RL11 proteins (eg. possess IgV-like domain). Although we managed to find members of the EE50 gene family that encode proteins with IgV-like domains in *Elephantid betaherpesvirus 1* and *Elephantid betaherpesvirus 5* (genus *Proboscivirus*), the encoded proteins lack a disulfide bond between β-strands C' and D, a characteristic mark of RL11 proteins, and likely were acquired by probosciviruses independently. We are aware that other viruses, not found in our study, may possess proteins with a similar protein structure. For example, NWM CMVs encode multiple proteins with IgV-like domains that shares some structural similarity with RL11 proteins (Martínez-Vicente et al. 2019, 2020; Pérez-Carmona et al. 2015), however the sequence similarity of these proteins was not sufficient to be identified in our screen. Thereby, we concluded that RL11 genes are specific to OWM and GA CMVs and they likely emerged after the divergence of NWM CMVs sometime between forty-two and twenty-nine million years ago (Kumar et al. 2022). Sequencing of CMV genomes infecting more basal primates such as loris, lemurs and tarsiers will show if the RL11 gene family was a more ancient acquisition which was lost in the NWM CMV lineage.

The high sequence diversity and short sequence length of the proteins encoded by RL11 genes make it challenging to establish phylogenetic relationships within the RL11 gene family with confidence, however we managed to produce consistent phylogenies from a combination of sequence-based and structure-aware methods. We showed that the RL11 gene family forms four distinct clades (RL11-α, RL11-β, RL11-γ and RL11-δ), with the RL11-β clade being unique to OWM CMVs. However, these phylogenies show low bootstrap support

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

for basal nodes and, in some cases, suffer from long-branch attraction (Bergsten 2005), and therefore should be interpreted with caution. One possible remedy for addressing long-branch attraction involves increasing the taxonomic sampling, which helps to break long branches and improves the estimation of the substitution model parameter. This problem can be addressed by increased sampling of RL11 genes from CMVs that infect other GA and OWM hosts such as orangutan, guenon, colobus, etc. Interestingly, several functional studies were conducted on the members of the RL11-α clade from *Macacine betaherpesvirus 3* (RL11A, RL11G and RL11T) demonstrating the IgG Fc-binding activity characteristic for the members of RL11-α clade from *Human betaherpesvirus 5* (RL11, RL12 and RL13) (Kolb et al. 2019; Otero et al. 2024; Taher et al. 2020). These findings clearly highlight the conservation of function of RL11-α genes across GA and OWM CMVs.

In our work, we defined the orthologous relationships between RL11 genes of OWM CMVs and GA CMVs. RL11 genes in these groups of CMVs are annotated in different ways. RL11 genes of OWM CMVs are named RL11A, RL11B – RL11T that reflects their belonging to the RL11 gene family and the order in which these genes are located in the CMV genome. On the other hand, for historical reasons two different types of names are used to refer to RL11 genes in GA CMVs. Five gene names start with RL (RL5A, RL6, RL11, RL12, and RL13) and nine gene names start with UL (UL1, UL4, UL5 – UL11) where numbers are used to indicate the order of the genes in the genome and letters are used to refer to the region where they were initially found. UL stands for the U_L region and RL stands for the terminal repeat of the U_L (that was an artefact due to genome rearrangements in a passaged virus strain; in the wild type CMVs all RL11 genes are found in the U_L). Our phylogenetic reconciliation analysis indicates that the RL11 gene family was shaped by a series of extensive duplication events

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

early in the evolution of OWM and GA CMVs with the largest number of duplication events occurring basally following the divergence from NWM CMVs. Losses of various RL11 genes happened more recently and usually in a lineage-specific manner. Therefore, the RL11 gene family represents a complex case of evolutionary relationships between family members where most RL11 genes from GA CMVs do not have one-to-one orthologs with OWM CMVs, apart from several exceptions like RL11 and RL11G, UL4 and RL11H (some OWM CMVs have two coorthologs – RL11H and RL11I), UL6 and RL11J, and probably UL7 and RL11P.

In complicated cases like this, it is particularly important to carefully choose appropriate terms to reflect the relationships between specific family members. For instance, many functional studies of RL11 genes were performed without a defined phylogeny of this gene family and therefore suffered from incorrect assignment of orthologous/paralogous relationships between family members. Here, we would like to address some of them. Firstly, our reconciliation analysis indicates that RL11G genes from OWM CMVs are orthologous to RL11 genes from GA CMVs and not to RL13 genes as was previously believed (Taher et al. 2020). This is interesting as previous studies showed that both RL13 and RL11G share a similar function of restricting the viral spread in fibroblasts (Schultz et al. 2020; Taher et al. 2020), suggesting that these genes might have independently acquired this function in GA (RL13) and OWM CMVs (RL11G). It is also possible that other genes in the RL11-α clade share a similar ancestral function, but to our knowledge functional analysis of the latter have not yet been investigated.

Secondly, we showed that RL11T from OWM CMVs have two coorthologs in GA CMVs, RL13 and RL12, where RL12 is likely an isoortholog (coortholog that retains the structure and

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

function of the ancestral gene after a duplication), while one-to-one orthologous relationship between RL12 and RL11T was previously assumed (Otero et al. 2024). Thirdly, the phylogenetic analysis shows that *Human betaherpesvirus 5*-specific genes RL5A and RL6 are coorthologs of the RL11A gene from OWM CMVs (but not necessarily isoorthologs), while before it was thought that the RL11A gene did not have an ortholog in GA CMVs (Kolb et al. 2019). Lastly, we would like to point out that according to our analysis, *Human betaherpesvirus 5*-specific gene UL1 is closely related to the UL10 gene where both genes likely emerged as the result of GA CMV-specific duplication event (Figure S6) and not as a duplication of RL11, RL12 or RL13 in the *Human betaherpesvirus 5* lineage as was previously speculated (Shikhagaie et al. 2012).

Although it was not the main focus of our study, we confirmed that CR1-β, -γ, and -δ genes in human adenoviruses and CR1-β genes in simian adenoviruses (genus *Mastadenovirus*) share noticeable sequence and structure similarity with RL11 genes. This finding is consistent with previously published data that human adenoviruses have genes potentially related to the RL11 gene family (Davison, Akter, et al. 2003). We showed that although CR1 genes form a distinct clade on a phylogenetic tree, CR1-γ and CR1-β genes encode proteins with IgV-like domains that share the topology and conserved disulfide bond with RL11 proteins. CR1-δ genes on the other hand likely lost the region corresponding to the C'' β-strand because of a deletion while the top part of the C' β-strand carrying a conserved cysteine was adapted into an α-helix. Also, the majority of the CR1 proteins carry two Ig-like domains (IgV-like and IgC-like) while proteins with one, two or three IgV-like domains are unique to human CR1-γ and human CR1-δ proteins. This could suggest that CR1-β genes with two Ig-like domains represent the ancestral form of CR1 proteins. We could not find

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

CR1-α genes in our screen suggesting that although these genes are related to CR1-β, -γ, and -δ, they share low sequence similarity with RL11 genes and likely require a lowered E-value/bitscore threshold to be found with RL11 probes.

Although the relationship between CR1 and RL11 gene families remains unclear, it is striking that members of two distantly related families of DNA viruses that infect primates possess a set of proteins with a common 3D structure. We can see three possible scenarios of how this could have happened. The first scenario is that the founder gene of the CR1 and RL11 gene families is the same gene that was independently acquired by CMVs and mastadenoviruses from a primate host. In this case we would expect that primates have a protein with the same IgV-like domain topology and conserved disulfide bond in the same place. There are primate proteins that possess the same IgV-like domain topology, e.g. CD244/2B4 (Velikovsky et al. 2007) or CD226/DNAM-1 (Wang et al. 2019). Similarly, to the CR1-β proteins, both CD244 and CD226 have an IgV domain followed by an IgC domain in their extracellular regions, but these IgV domains have a disulfide bond in a different place. The extracellular region of the human CD229/SLAMF3/LY9 on the other hand contains four Ig domains (IgV1, IgC1, IgV2 and IgC2) where IgV1 also shares the topology but not the disulfide bond with the RL11D and CR1 domains (Varadi et al. 2024). The second hypothesis is that the founder gene was acquired from a host only by one virus genus. In this virus genus, the gene underwent a rapid evolution and acquired a disulfide bond between β-strands C' and D which was later co-opted by another virus genus probably as a result of coinfection. The third scenario is that both CMVs and adenoviruses acquired two different genes with IgV-like fold from their hosts and adopted the same topology with a conserved disulfide bond through convergent evolution.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

Each of these hypotheses relies on the virus acquisition of a host gene and it is known that DNA viruses can co-opt mammalian genes via retrotransposition (Fixsen et al. 2022). Using our approach which was focused on understanding the evolution of the RL11 gene family in CMVs, we did not find any RL11-like matches in the mammalian hosts that satisfied our E-value/bitscore threshold. Future work using a protein-centred rather than a genome-centred screening approach might help to elucidate the mammalian origin of the RL11 gene family. However, we should point out that considering the high diversity of RL11 genes and the already limited bootstrap support of basal phylogenetic relationships, deciphering with confidence these deeper relationships will be challenging.

The extensive species-specific evolution of the RL11 gene family in primates during co-evolution of their respective CMVs implies that these genes play key roles by interacting with components of the host immune system undergoing interdependent selection. Interestingly, not only is this gene family variable in sequence and arrangement between species, but the genes in the family are also the most highly variable between virus isolates within the same species. Many fall into ten or more genotypes, suggesting significant within-host evolution has also occurred since species divergence (Sijmons et al. 2015; Suárez et al. 2019). Identification of orthologs across different CMVs from this study may now facilitate dissection of functional homology, and the underlying reason for this extensive lineage-specific gene diversification.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

Data availability

All data generated in this study including Python scripts, raw DIGS results, multiple sequence alignments, maximum likelihood phylogenetic trees, predicted protein structures and predictions of functional regions are available at https://github.com/ulad-litvin/cmv_rl11_evolutionary_dynamics.

Acknowledgements

We thank Professor Andrew Davison for useful feedback during the project and Dr Rob Gifford with help applying DIGS.

Funding

U.L. was supported by a PhD studentship from the Darwin Trust of Edinburgh. J.H. acknowledge funding from the UK Medical Research Council (MRC, MC_UU_12014/12 and MC_UU_00034/5). R.S. was funded by the Wellcome Trust (226615/Z/22/Z).

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

References

Atalay, R., Zimmermann, A., Wagner, M., Borst, E., Benz, C., Messerle, M., & Hengel, H. (2002). 'Identification and Expression of Human Cytomegalovirus Transcription Units Coding for Two Distinct Fcγ Receptor Homologs', *Journal of Virology*, 76/17: 8596–608. American Society for Microbiology. DOI: 10.1128/jvi.76.17.8596-8608.2002

Bergsten, J. (2005). 'A review of long-branch attraction', *Cladistics*, 21/2: 163–93. DOI: 10.1111/j.1096-0031.2005.00059.x

Blanco-Melo, D., Campbell, M. A., Zhu, H., Dennis, T. P. W., Modha, S., Lytras, S., Hughes, J., et al. (2023). 'A novel approach to exploring the dark genome and its application to mapping of the vertebrate virus "fossil record"'. bioRxiv. DOI: 10.1101/2023.10.17.562709

Blewett, E. L., Sherrod, C. J., Texier, J. R., Conrad, T. M., & Dittmer, D. P. (2015). 'Complete Genome Sequences of Mandrillus leucophaeus and Papio ursinus Cytomegaloviruses', *Genome Announcements*, 3/4: 10.1128/genomea.00781-15. American Society for Microbiology. DOI: 10.1128/genomea.00781-15

Brito, A. F., Baele, G., Nahata, K. D., Grubaugh, N. D., & Pinney, J. W. (2021). 'Intrahost speciations and host switches played an important role in the evolution of herpesviruses', *Virus Evolution*, 7/1: veab025. DOI: 10.1093/ve/veab025

Bruno, L., Cortese, M., Monda, G., Gentile, M., Calò, S., Schiavetti, F., Zedda, L., et al. (2016). 'Human cytomegalovirus pUL10 interacts with leukocytes and impairs TCR-mediated T-cell activation', *Immunology & Cell Biology*, 94/9: 849–60. DOI: 10.1038/icb.2016.49

Cagliani, R., Forni, D., Mozzi, A., & Sironi, M. (2020). 'Evolution and Genetic Diversity of Primate Cytomegaloviruses', *Microorganisms*, 8/5: 624. Multidisciplinary Digital Publishing Institute. DOI: 10.3390/microorganisms8050624

Cortese, M., Calò, S., D'Aurizio, R., Lilja, A., Pacchiani, N., & Merola, M. (2012). 'Recombinant Human Cytomegalovirus (HCMV) RL13 Binds Human Immunoglobulin G Fc', *PLOS ONE*, 7/11: e50166. Public Library of Science. DOI: 10.1371/journal.pone.0050166

Davison, A. J., Akter, P., Cunningham, C., Dolan, A., Addison, C., Dargan, D. J., Hassan-Walker, A. F., et al. (2003). 'Homology between the human cytomegalovirus RL11 gene family and human adenovirus E3 genes', *Journal of General Virology*, 84/3: 657–63. Microbiology Society,. DOI: 10.1099/vir.0.18856-0

Davison, A. J., Dolan, A., Akter, P., Addison, C., Dargan, D. J., Alcendor, D. J., McGeoch, D. J., et al. (2003). 'The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genomeFN1', *Journal of General Virology*, 84/1: 17–28. Microbiology Society,. DOI: 10.1099/vir.0.18606-0

Davison, A. J., Holton, M., Dolan, A., Dargan, D. J., Gatherer, D., & Hayward, G. S. (2013). 'Comparative Genomics of Primate Cytomegaloviruses'. Reddehase M. J. (ed.) *Cytomegaloviruses: From Molecular Pathogenesis to Intervention*, Vols 1-2, Vol. 1, p. xxiv + 464. Caister Academic Press: Norfolk, UK.

Engel, P., Pérez-Carmona, N., Albà, M. M., Robertson, K., Ghazal, P., & Angulo, A. (2011). 'Human cytomegalovirus UL7, a homologue of the SLAM-family receptor CD229,

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

impairs cytokine production', *Immunology & Cell Biology*, 89/7: 753–66. DOI: 10.1038/icb.2011.55

Fixsen, S. M., Cone, K. R., Goldstein, S. A., Sasani, T. A., Quinlan, A. R., Rothenburg, S., & Elde, N. C. (2022). 'Poxviruses capture host genes by LINE-1 retrotransposition', (K. Kirkegaard, G. H. Perry, E. V. Koonin, & D. Walsh, Eds)*eLife*, 11: e63332. eLife Sciences Publications, Ltd. DOI: 10.7554/eLife.63332

Gabaev, I., Steinbrück, L., Pokoyski, C., Pich, A., Stanton, R. J., Schwinzer, R., Schulz, T. F., et al. (2011). 'The Human Cytomegalovirus UL11 Protein Interacts with the Receptor Tyrosine Phosphatase CD45, Resulting in Functional Paralysis of T Cells', *PLOS Pathogens*, 7/12: e1002432. Public Library of Science. DOI: 10.1371/journal.ppat.1002432

Griffiths, P., & Reeves, M. (2021). 'Pathogenesis of human cytomegalovirus in the immunocompromised host', *Nature Reviews Microbiology*, 19/12: 759–73. Nature Publishing Group. DOI: 10.1038/s41579-021-00582-z

Gupta, R., & Brunak, S. (2002). 'Prediction of glycosylation across the human proteome and the correlation to protein function', *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 310–22.

Hitt, M., Bett, A. J., Addison, C. L., Prevec, L., & Graham, F. L. (1995). '[2] Techniques for human adenovirus vector construction and characterization'. Adolph K. W. (ed.) *Methods in Molecular Genetics*, Viral Gene Techniques, Vol. 7, pp. 13–30. Academic Press. DOI: 10.1016/S1067-2389(06)80034-8

de la Hoz, R. E., Stephens, G., & Sherlock, C. (2002). 'Diagnosis and treatment approaches of CMV infections in adult patients', *Journal of Clinical Virology*, 25: 1–12. DOI: 10.1016/S1386-6532(02)00091-4

Jacobs, S. C., Davison, A. J., Carr, S., Bennett, A. M., Phillpotts, R., & Wilkinson, G. W. G. (2004). 'Characterization and manipulation of the human adenovirus 4 genome', *Journal of General Virology*, 85/11: 3361–6. Microbiology Society,. DOI: 10.1099/vir.0.80386-0

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., et al. (2014). 'InterProScan 5: genome-scale protein function classification', *Bioinformatics*, 30/9: 1236–40. DOI: 10.1093/bioinformatics/btu031

Katoh, K., & Standley, D. M. (2013). 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30/4: 772–80. DOI: 10.1093/molbev/mst010

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., et al. (2023). 'Fast and accurate protein structure search with Foldseek', *Nature Biotechnology*, 1–4. Nature Publishing Group. DOI: 10.1038/s41587-023-01773-0

Kolb, P., Sijmons, S., McArdle, M. R., Taher, H., Womack, J., Hughes, C., Ventura, A., et al. (2019). 'Identification and Functional Characterization of a Novel Fc Gamma-Binding Glycoprotein in Rhesus Cytomegalovirus', *Journal of Virology*, 93/4: 10.1128/jvi.02077-18. American Society for Microbiology. DOI: 10.1128/jvi.02077-18

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., et al. (2022). 'TimeTree 5: An Expanded Resource for Species Divergence Times', *Molecular Biology and Evolution*, 39/8: msac174. DOI: 10.1093/molbev/msac174

Laskowski, R. A., Jabłońska, J., Pravda, L., Vařeková, R. S., & Thornton, J. M. (2018). 'PDBsum: Structural summaries of PDB entries', *Protein Science*, 27/1: 129–34. DOI: 10.1002/pro.3289

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G., & Smith, D. B. (2018). ‘Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)’, *Nucleic Acids Research*, 46/D1: D708–17. DOI: 10.1093/nar/gkx932

Lemoine, F., Domelevo Entfellner, J.-B., Wilkinson, E., Correia, D., Dávila Felipe, M., De Oliveira, T., & Gascuel, O. (2018). ‘Renewing Felsenstein’s phylogenetic bootstrap in the era of big data’, *Nature*, 556/7702: 452–6. Nature Publishing Group. DOI: 10.1038/s41586-018-0043-0

Lilley, B. N., Ploegh, H. L., & Tirabassi, R. S. (2001). ‘Human Cytomegalovirus Open Reading Frame TRL11/IRL11 Encodes an Immunoglobulin G Fc-Binding Protein’, *Journal of Virology*, 75/22: 11218–21. American Society for Microbiology. DOI: 10.1128/jvi.75.22.11218-11221.2001

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., et al. (2023). ‘Evolutionary-scale prediction of atomic-level protein structure with a language model’, *Science*, 379/6637: 1123–30. American Association for the Advancement of Science. DOI: 10.1126/science.ade2574

Marsh, A. K., Willer, D. O., Ambagala, A. P. N., Dzamba, M., Chan, J. K., Pilon, R., Fournier, J., et al. (2011). ‘Genomic Sequencing and Characterization of Cynomolgus Macaque Cytomegalovirus’, *Journal of Virology*, 85/24: 12995–3009. American Society for Microbiology. DOI: 10.1128/jvi.05840-11

Martínez-Vicente, P., Farré, D., Engel, P., & Angulo, A. (2020). ‘Divergent Traits and Ligand-Binding Properties of the Cytomegalovirus CD48 Gene Family’, *Viruses*, 12/8: 813. Multidisciplinary Digital Publishing Institute. DOI: 10.3390/v12080813

Martínez-Vicente, P., Farré, D., Sánchez, C., Alcamí, A., Engel, P., & Angulo, A. (2019). ‘Subversion of natural killer cell responses by a cytomegalovirus-encoded soluble CD48 decoy receptor’, *PLOS Pathogens*, 15/4: e1007658. Public Library of Science. DOI: 10.1371/journal.ppat.1007658

McGeoch, D. J., Rixon, F. J., & Davison, A. J. (2006). ‘Topics in herpesvirus genomics and evolution’, *Virus Research*, Comparative Genomics and Evolution of Complex Viruses, 117/1: 90–104. DOI: 10.1016/j.virusres.2006.01.002

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). ‘IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era’, *Molecular Biology and Evolution*, 37/5: 1530–4. DOI: 10.1093/molbev/msaa015

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ‘ColabFold: making protein folding accessible to all’, *Nature Methods*, 19/6: 679–82. Nature Publishing Group. DOI: 10.1038/s41592-022-01488-1

Morel, B., Kozlov, A. M., Stamatakis, A., & Szöllősi, G. J. (2020). ‘GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss’, *Molecular Biology and Evolution*, 37/9: 2763–74. DOI: 10.1093/molbev/msaa141

Murthy, S., O’Brien, K., Agbor, A., Angedakin, S., Arandjelovic, M., Ayimisin, E. A., Bailey, E., et al. (2019). ‘Cytomegalovirus distribution and evolution in hominines’, *Virus Evolution*, 5/2: vez015. DOI: 10.1093/ve/vez015

Otero, C. E., Petkova, S., Ebermann, M., Taher, H., John, N., Hoffmann, K., Davalos, A., et al. (2024). ‘Rhesus Cytomegalovirus-encoded Fcγ-binding glycoproteins facilitate viral evasion from IgG-mediated humoral immunity’. bioRxiv. DOI: 10.1101/2024.02.27.582371

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

Pérez-Carmona, N., Farré, D., Martínez-Vicente, P., Terhorst, C., Engel, P., & Angulo, A. (2015). ‘Signaling Lymphocytic Activation Molecule Family Receptor Homologs in New World Monkey Cytomegaloviruses’, *Journal of Virology*, 89/22: 11323–36. American Society for Microbiology. DOI: 10.1128/jvi.01296-15

Pérez-Carmona, N., Martínez-Vicente, P., Farré, D., Gabaev, I., Messerle, M., Engel, P., & Angulo, A. (2018). ‘A Prominent Role of the Human Cytomegalovirus UL8 Glycoprotein in Restraining Proinflammatory Cytokine Production by Myeloid Cells at Late Times during Infection’, *Journal of Virology*, 92/9: 10.1128/jvi.02229-17. American Society for Microbiology. DOI: 10.1128/jvi.02229-17

Schultz, E. P., Lanchy, J.-M., Day, L. Z., Yu, Q., Peterson, C., Preece, J., & Ryckman, B. J. (2020). ‘Specialization for Cell-Free or Cell-to-Cell Spread of BAC-Cloned Human Cytomegalovirus Strains Is Determined by Factors beyond the UL128-131 and RL13 Loci’, *Journal of Virology*, 94/13: 10.1128/jvi.00034-20. American Society for Microbiology. DOI: 10.1128/jvi.00034-20

Sekulin, K., Görzer, I., Heiss-Czedik, D., & Puchhammer-Stöckl, E. (2007). ‘Analysis of the variability of CMV strains in the RL11D domain of the RL11 multigene family’, *Virus Genes*, 35/3: 577–83. DOI: 10.1007/s11262-007-0158-0

Shikhagaie, M., Mercé-Maldonado, E., Isern, E., Muntasell, A., Albà, M. M., López-Botet, M., Hengel, H., et al. (2012). ‘The Human Cytomegalovirus-Specific UL1 Gene Encodes a Late-Phase Glycoprotein Incorporated in the Virion Envelope’, *Journal of Virology*, 86/8: 4091–101. American Society for Microbiology. DOI: 10.1128/jvi.06291-11

Sijmons, S., Thys, K., Mbong Ngwese, M., Van Damme, E., Dvorak, J., Van Loock, M., Li, G., et al. (2015). ‘High-Throughput Analysis of Human Cytomegalovirus Genome Diversity Highlights the Widespread Occurrence of Gene-Disrupting Mutations and Pervasive Recombination’, *Journal of Virology*, 89/15: 7673–95. American Society for Microbiology. DOI: 10.1128/jvi.00578-15

Singh, G., Robinson, C. M., Dehghan, S., Jones, M. S., Dyer, D. W., Seto, D., & Chodosh, J. (2013). ‘Homologous Recombination in E3 Genes of Human Adenovirus Species D’, *Journal of Virology*, 87/22: 12481–8. American Society for Microbiology. DOI: 10.1128/jvi.01927-13

Steentoft, C., Vakhrushev, S. Y., Joshi, H. J., Kong, Y., Vester-Christensen, M. B., Schjoldager, K. T. G., Lavrsen, K., et al. (2013). ‘Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology’, *The EMBO Journal*, 32/10: 1478–88. John Wiley & Sons, Ltd. DOI: 10.1038/emboj.2013.79

Steenwyk, J. L., Iii, T. J. B., Li, Y., Shen, X.-X., & Rokas, A. (2020). ‘ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference’, *PLOS Biology*, 18/12: e3001007. Public Library of Science. DOI: 10.1371/journal.pbio.3001007

Suárez, N. M., Wilkie, G. S., Hage, E., Camiolo, S., Holton, M., Hughes, J., Maabar, M., et al. (2019). ‘Human Cytomegalovirus Genomes Sequenced Directly From Clinical Material: Variation, Multiple-Strain Infection, Recombination, and Gene Loss’, *The Journal of Infectious Diseases*, 220/5: 781–91. DOI: 10.1093/infdis/jiz208

Taher, H., Mahyari, E., Kreklywich, C., Uebelhoer, L. S., McArdle, M. R., Moström, M. J., Bhusari, A., et al. (2020). ‘In vitro and in vivo characterization of a recombinant rhesus cytomegalovirus containing a complete genome’, *PLOS Pathogens*, 16/11: e1008666. Public Library of Science. DOI: 10.1371/journal.ppat.1008666

Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., Gíslason, M. H., Pihl, S. I., Tsirigos, K. D., Winther, O., et al. (2022). ‘SignalP 6.0 predicts all five types of signal peptides using

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

protein language models', *Nature Biotechnology*, 40/7: 1023–5. Nature Publishing Group. DOI: 10.1038/s41587-021-01156-3

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., et al. (2024). 'AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences', *Nucleic Acids Research*, 52/D1: D368–75. DOI: 10.1093/nar/gkad1011

Velikovsky, C. A., Deng, L., Chlewicki, L. K., Fernández, M. M., Kumar, V., & Mariuzza, R. A. (2007). 'Structure of Natural Killer Receptor 2B4 Bound to CD48 Reveals Basis for Heterophilic Recognition in Signaling Lymphocyte Activation Molecule Family', *Immunity*, 27/4: 572–84. DOI: 10.1016/j.immuni.2007.08.019

Vlachava, V.-M., Seirafian, S., Fielding, C. A., Kollnberger, S., Aicheler, R. J., Hughes, J., Baker, A., et al. (2023). 'HCMV-secreted glycoprotein gpUL4 inhibits TRAIL-mediated apoptosis and NK cell activation', *Proceedings of the National Academy of Sciences*, 120/49: e2309077120. Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.2309077120

Wang, H., Qi, J., Zhang, S., Li, Y., Tan, S., & Gao, G. F. (2019). 'Binding mode of the side-by-side two-IgV molecule CD226/DNAM-1 to its ligand CD155/Necl-5', *Proceedings of the National Academy of Sciences*, 116/3: 988–96. Proceedings of the National Academy of Sciences. DOI: 10.1073/pnas.1815716116

Zuhair, M., Smit, G. S. A., Wallis, G., Jabbar, F., Smith, C., Devleesschauwer, B., & Griffiths, P. (2019). 'Estimation of the worldwide seroprevalence of cytomegalovirus: A systematic review and meta-analysis', *Reviews in Medical Virology*, 29/3: e2034. DOI: 10.1002/rmv.2034

Figure 1:

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

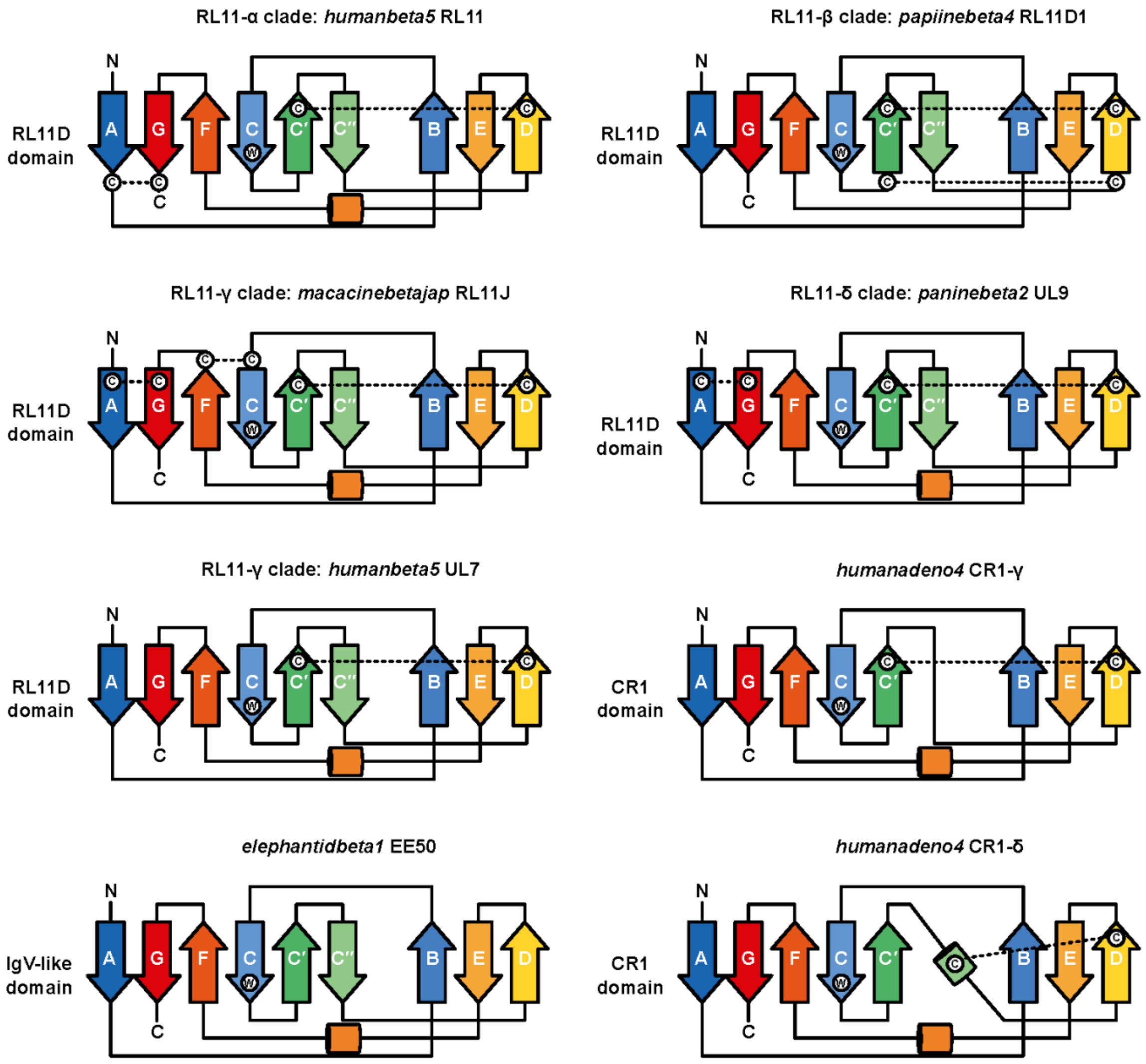

RL11-α clade: *humanbeta5* RL11
RL11-β clade: *papiinebeta4* RL11D1
RL11-γ clade: *macacinebetajap* RL11J
RL11-δ clade: *paninebeta2* UL9
RL11-γ clade: *humanbeta5* UL7
humanadeno4 CR1-γ
elephantidbeta1 EE50
humanadeno4 CR1-δ
RL11D domain
CR1 domain
IgV-like domain
N
C
A
G
F
C
C′
C″
B
E
D
W

Figure 2:

ACCEPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024

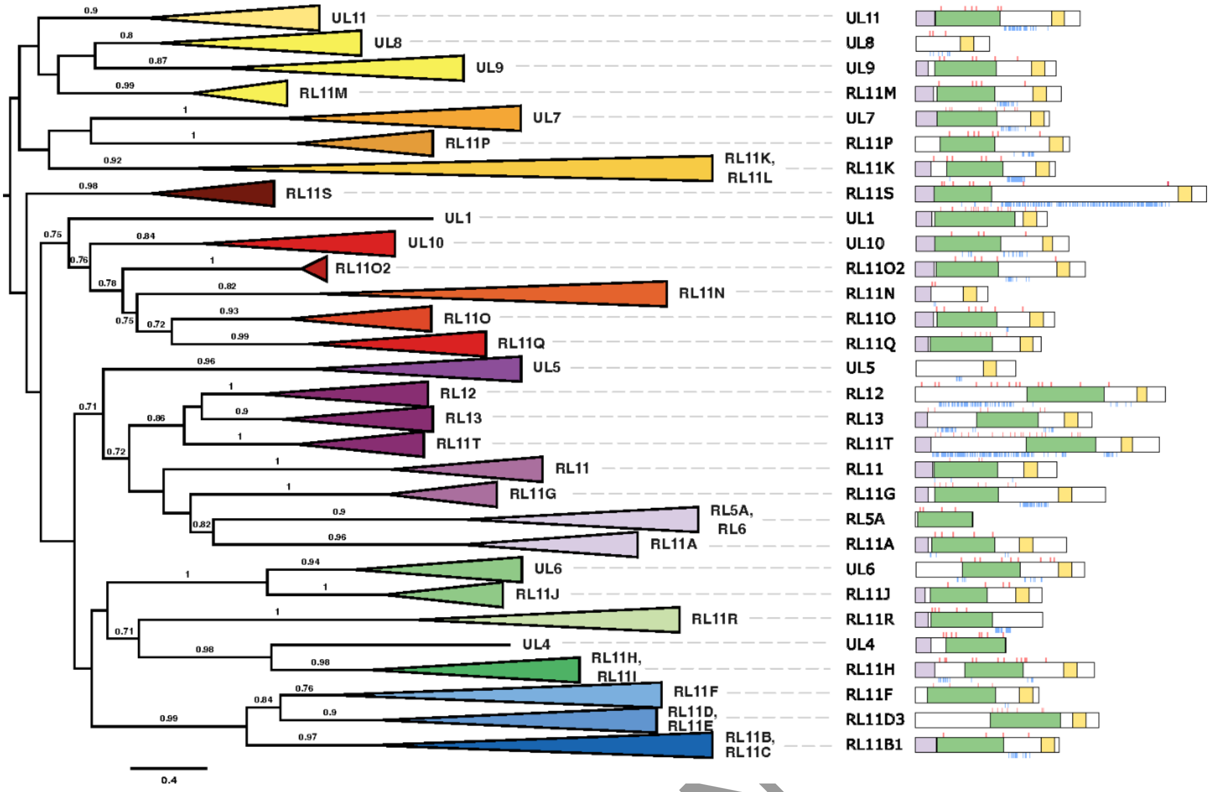
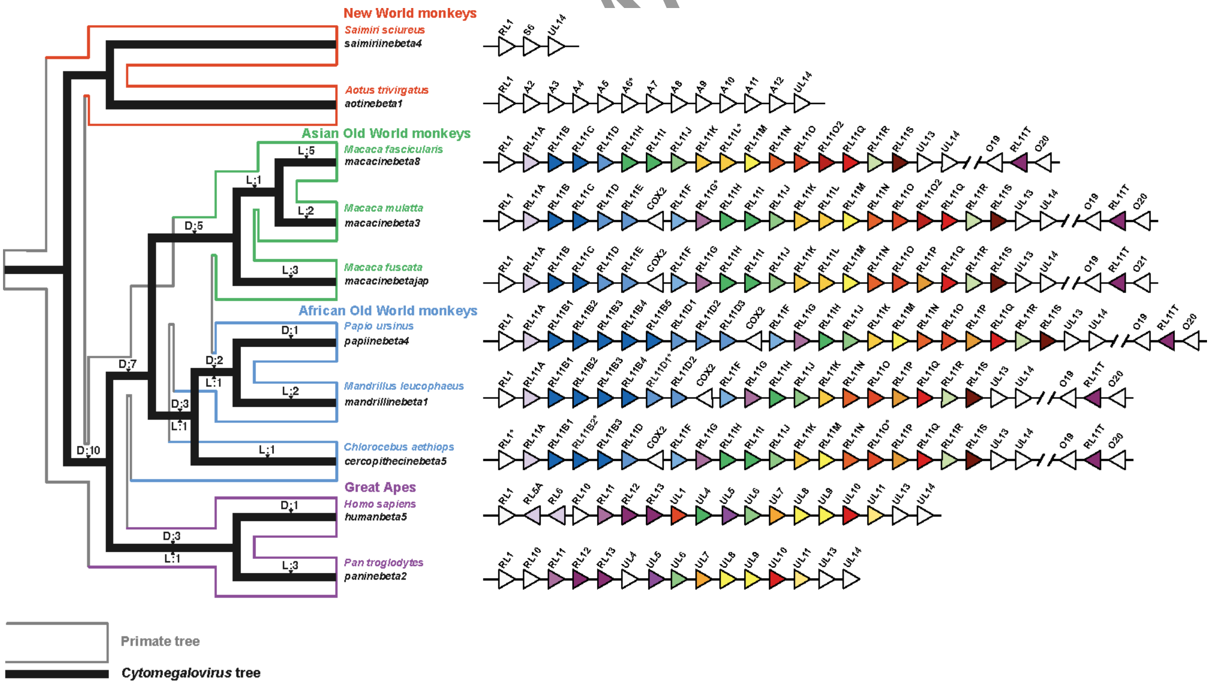


Figure 3:


New World monkeys
Saimiri sciureus
saimiriinebeta4
Aotus trivirgatus
aotinebeta1
Asian Old World monkeys
Macaca fascicularis
macacinebeta8
Macaca mulatta
macacinebeta3
Macaca fuscata
macacinebetajap
African Old World monkeys
Papio ursinus
papiinebeta4
Mandrillus leucophaeus
mandrillinebeta1
Chlorocebus aethiops
cercopithecinebeta5
Great Apes
Homo sapiens
humanbeta5
Pan troglodytes
paninebeta2
D:10
D:7
D:5
L:1
L:5
L:2
L:3
D:1
D:2
L:1
D:3
L:1
L:2
L:1
D:3
L:1
D:1
L:3
Primate tree
Cytomegalovirus tree
RL1 S6 UL14
RL1 A2 A3 A4 A5 A6* A7 A8 A9 A10 A11 A12 UL14
RL1 RL11A RL11B RL11C RL11D RL11H RL11I RL11J RL11K RL11L* RL11M RL11N RL11O RL11O2 RL11Q RL11R RL11S UL13 UL14 O19 RL11T O20
RL1 RL11A RL11B RL11C RL11D RL11E COX2 RL11F RL11G* RL11H RL11I RL11J RL11K RL11L RL11M RL11N RL11O RL11O2 RL11Q RL11R RL11S UL13 UL14 O19 RL11T O20
RL1 RL11A RL11B RL11C RL11D RL11E COX2 RL11F RL11G RL11H RL11I RL11J RL11K RL11L RL11M RL11N RL11O RL11P RL11Q RL11R RL11S UL13 UL14 O19 RL11T O21
RL1 RL11A RL11B1 RL11B2 RL11B3 RL11B4 RL11B5 RL11D1 RL11D2 RL11D3 COX2 RL11F RL11G RL11H RL11J RL11K RL11M RL11N RL11O RL11P RL11Q RL11R RL11S UL13 UL14 O19 RL11T O20
RL1 RL11A RL11B1 RL11B2 RL11B3 RL11B4 RL11D1* RL11D2 COX2 RL11F RL11G RL11H RL11I RL11J RL11K RL11N RL11O RL11P RL11Q RL11R RL11S UL13 UL14 O19 RL11T O20
RL1* RL11A RL11B1 RL11B2* RL11B3 RL11D COX2 RL11F RL11G RL11H RL11I RL11J RL11K RL11M RL11N RL11O* RL11P RL11Q RL11R RL11S UL13 UL14 O19 RL11T O20
RL1 RL5A RL6 RL10 RL11 RL12 RL13 UL1 UL4 UL5 UL6 UL7 UL8 UL9 UL10 UL11 UL13 UL14
RL1 RL10 RL11 RL12 RL13 UL4 UL5 UL6 UL7 UL8 UL9 UL10 UL11 UL13 UL14

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veae066/7741204 by guest on 04 September 2024