AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

GENETIC DIVERSITY AND EVOLUTION

Cell-Derived Viral Genes Evolve under Stronger Purifying Selection in Rhadinoviruses

Amr Aswad,[a] Aris Katzourakis[a]

[a]Zoology Department, University of Oxford, Oxford, Oxfordshire, United Kingdom

ABSTRACT Like many other large double-stranded DNA (dsDNA) viruses, herpesviruses are known to capture host genes to evade host defenses. Little is known about the detailed natural history of such genes, nor do we fully understand their evolutionary dynamics. A major obstacle is that they are often highly divergent, maintaining very low sequence similarity to host homologs. Here we use the herpesvirus genus *Rhadinovirus* as a model system to develop an analytical approach that combines complementary evolutionary and bioinformatic techniques, offering results that are both detailed and robust for a range of genes. Using a systematic phylogenetic strategy, we identify the original host lineage of viral genes with high confidence. We show that although host immunomodulatory genes evolve rapidly compared to other host genes, they undergo a clear increase in purifying selection once captured by a virus. To characterize this shift in detail, we developed a novel technique to identify changes in selection pressure that can be attributable to particular domains. These findings will inform us on how viruses develop strategies to evade the immune system, and our synthesis of techniques can be reapplied to other viruses or biological systems with similar analytical challenges.

IMPORTANCE Viruses and hosts have been shown to capture genes from one another as part of the evolutionary arms race. Such genes offer a natural experiment on the effects of evolutionary pressure, since the same gene exists in vastly different selective environments. However, sequences of viral homologs often bear little similarity to the original sequence, complicating the reconstruction of their shared evolutionary history with host counterparts. In this study, we use a genus of herpesviruses as a model system to comprehensively investigate the evolution of host-derived viral genes, using a synthesis of genomics, phylogenetics, selection analysis, and nucleotide and amino acid modeling.

KEYWORDS host-derived viral genes, paleovirology, rhadinoviruses, virus evolution

The *Herpesviridae* are a diverse family of large double-stranded DNA (dsDNA) viruses that infect mammals, birds, and reptiles. Among the three subfamilies (*Alpha-*, *Beta-*, and *Gammaherpesvirinae*), the *Gammaherpesvirinae* form a distinct phylogenetic lineage, and many members share a number of genes that are unique to the subfamily. Furthermore, many gammaherpesviruses infect lymphocytes and are known to cause lymphoproliferative disorders. The *Gammaherpesvirinae* subfamily is further split into the four genera *Lymphocryptovirus*, *Rhadinovirus*, *Macavirus*, and *Percavirus*, based on phylogenetically distinct grouping. All gammaherpesviruses infect mammals, and they include two of the eight known human herpesviruses. The *Lymphocryptovirus* type species is *Human herpesvirus 4* (also known as *Epstein-Barr virus*), while the Kaposi's sarcoma-causing *Human herpesvirus 8* (HHV8) is the type species of rhadinoviruses.

Like other herpesviruses (and indeed many large dsDNA viruses), rhadinoviruses can adaptively capture host genes that increase their fitness through horizontal gene transfer (HGT). As well as "core blocks" of genes that are conserved across herpesvi-

Received 2 March 2018 **Accepted** 1 June 2018

Accepted manuscript posted online 11 July 2018

Citation Aswad A, Katzourakis A. 2018. Cell-derived viral genes evolve under stronger purifying selection in rhadinoviruses. J Virol 92:e00359-18. https://doi.org/10.1128/JVI.00359-18.

Editor Jae U. Jung, University of Southern California

Address correspondence to Aris Katzourakis, aris.katzourakis@zoo.ox.ac.uk.

ruses, lineage-specific genes include captured cellular immunomodulatory genes, such as those encoding cytokines, chemokines, and interferon regulatory factors (1–3). These genes evolve at a particularly high rate in vertebrates, driven by the selection pressure imposed by viruses.

The viral evolutionary strategy of cellular gene capture is so successful that some genes, such as interleukin 10, have repeatedly been captured multiple times independently in different viral groups (3, 4). Rhadinoviruses are an ideal study group for systematically reconstructing the evolutionary history of cell-derived genes, because they are known to frequently capture host genes and there are seven well-annotated genomes with lineage-specific cell-derived open reading frames (ORFs) (5). Moreover, the genus includes HHV8, the etiological agent of Kaposi's sarcoma, for which ORFs with sequence similarity to human genes have already been identified (2).

Gene flow between viruses and their hosts occurs in both directions, with numerous examples of viral genomes becoming heritably integrated into the genomes of their hosts. Known as endogenous viral elements (EVEs) (6–9), they are usually nonfunctional relics but can play a beneficial role in the host genome, including as antiviral defense genes, known as EVE-derived immunity genes (EDIs) (6, 10, 11). The capture of these EDIs is a mechanism employed by hosts as part of the evolutionary arms race to defend against viruses, just as viruses capture cellular genes to evade host immunity.

In order to explore the evolutionary dynamics of gene capture by rhadinoviruses, we use a combination of comparative genomics and phylogenetic reconstruction. We take a genome-wide approach to studying host-derived rhadinoviral genes in the same way that EDIs can be identified and investigated through the study of EVEs in paleovirology. In a sense, cell-derived genes are the conceptual counterpart to EVEs, i.e., a kind of "endogenous host element" in viral genomes.

The results of the present study conclusively demonstrate the mammalian origin of captured rhadinoviral genes, identifying the specific donor group for seven of them. We investigate the history of evolutionary pressure imposed on captured genes and develop an implementation of existing maximum likelihood selection analyses to reveal a detailed picture of the evolutionary dynamics within certain cell-derived genes. Our study reveals the remarkable finding that host-derived genes in rhadinoviruses are under stronger purifying selection (i.e., purging of deleterious mutations) than that in hosts, and we interpret this change as a means to maintain the original function of the gene product. This is further reflected in the fact that amino acid sequence similarity, though low, is maintained despite a dramatic shift in nucleotide composition away from that of the original gene. This is consistent with the long-standing observation that herpesvirus genes maintain a common functionality despite a shift in nucleotide composition (e.g., see reference 12, published over 30 years ago). Moreover, we identify evidence of this functional preservation in the predicted proteins that are structurally congruent to their cellular predecessors despite low sequence similarity. Through comparative genomic analysis of rhadinoviruses under an evolutionary framework, our study reveals that a shift toward stronger purifying selection is a common evolutionary mechanism underlying the capture of some cellular genes.

RESULTS

We constructed a whole-genome alignment of seven rhadinoviruses, using Mauve (13), to structure the investigation according to the locus of each putative gene capture (Fig. 1). We focused on genes outside the conserved core blocks, which we know to be shared across all herpesviruses, and arbitrarily designated the remaining regions variable regions A to C (Fig. 1). Each gene was matched to its likely homolog in other rhadinoviruses by searching for sequence similarity using BLASTp, with a minimum alignment score threshold of 50. Sequence-similar rhadinoviral genes were considered syntenic if they shared similar flanking genes in the equivalent region of the genome. BLAST searching was also used for each gene to identify similarity to host sequences, which is an indication that the viral sequence could have been derived by HGT. We excluded genes that were too divergent to align to similar host genes with confidence,

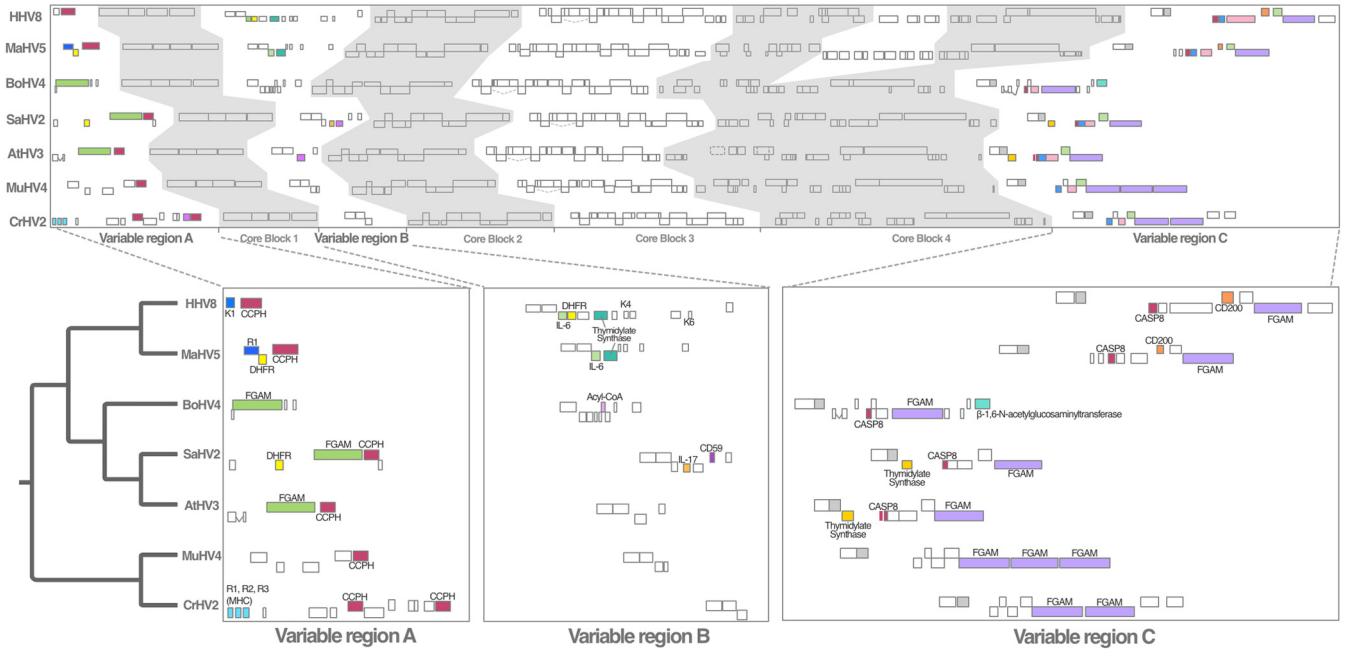

FIG 1 Whole-genome alignment of rhadinoviruses. The figure depicts a scaled cartoon of an alignment of seven rhadinoviral genomes, with boxes representing ORFs according to the GenBank annotation. Both lineage-specific (variable) and core herpesvirus blocks are indicated by the labels below the alignment and are highlighted in gray or white in an alternating fashion. The variable regions are further exploded for clarity, highlighting the genes that were analyzed in this study. The cladogram indicates the phylogenetic relationships between viruses. The region in HHV8 and MaHV5 corresponding to core block 4 includes a stretch of lineage-specific genes that is not part of the set of conserved genes. Horizontally acquired genes investigated in this analysis are indicated in color and annotated in the zoomed-in panels below the alignment. Virus abbreviations are as follows: HHV8, *Human herpesvirus 8* (also known as Kaposi's sarcoma-associated herpesvirus [KSHV]); MaHV5, *Macacine gammaherpesvirus 5*; BoHV4, *Bovine gammaherpesvirus 4*; SaHV2, *Saimiriine gammaherpesvirus 2*; AtHV3, *Ateline gammaherpesvirus 3*; MuHV4, *Murid gammaherpesvirus 4*; and CrHV2, *Cricetid gammaherpesvirus 2* (also known as *Rodent herpesvirus Peru* [RHVP]).

i.e., cases where homologous sites could be hypothesized for fewer than 50 amino acids. This resulted in a short list of 13 genes that we considered for further phylogenetic and evolutionary analyses.

MHC, interleukin-6 (IL-6), and IL-17 genes. The first three genes in *Cricetid gammaherpesvirus 2* (CrHV2), designated R1, R2, and R3, show sequence similarity to major histocompatibility complex (MHC) genes (Fig. 1) and were aligned in the initial characterization of the viral genome with human HLA-A2 and mouse H-2K^b (14). In our analysis of these genes, rather than treating them as separate genes, we determined that they are in fact similar to different regions of the MHC gene in a collinear manner. This lead us to the conclusion that they were derived from the capture of a single gene, and it was possible to align a concatenation of all three ORFs to a range of MHC homologs. We reconstructed a phylogeny from a diverse set of mammalian hosts, revealing that CrHV2 R1 to R3 are most closely related to the mouse and rat homologs (Fig. 2A). Along with the fact that CrHV2 was isolated from the pygmy rice rat (*Oligoryzomys microtis*), the phylogenetic similarity to mouse and rat (*Mus musculus* and *Rattus norvegicus*, respectively) is consistent with acquisition from a rodent host (Fig. 2A). However, none of the mammalian MHC genes we identified possesses an exon structure that corresponds to the three CrHV2 ORFs, suggesting that the rodent in question has not been sequenced and may or may not be extant.

Interleukins are a large group of cytokines, among which genes from several different families have been captured by rhadinoviruses. IL-17 is a family of 6 types of interleukin involved in the inflammatory response (IL-17A to -F) (15). *Saimiriine gammaherpesvirus 2* (SaHV2) acquired a copy of the IL-17A gene, and the fact that this gene is not present in *Ateline gammaherpesvirus 3* (AtHV3) suggests that the capture event occurred after their speciation. Phylogenetic analysis is consistent with this hypothesis (Fig. 3A), since despite the long branch of the SaHV2 IL-17A gene, the gene groups robustly with the IL-17 genes of squirrel monkey (*Saimiri boliviensis*) and marmoset

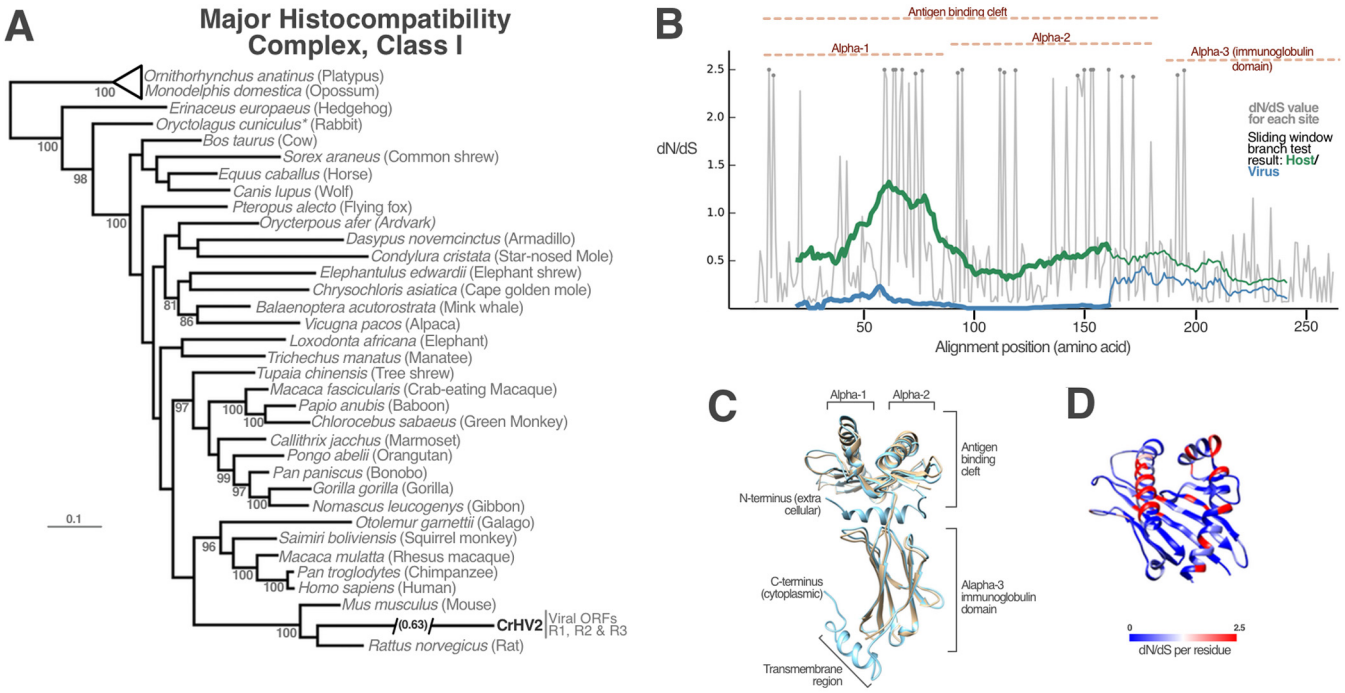

FIG 2 Capture of a major histocompatibility complex (MHC) gene by CrHV2. (A) Bayesian phylogenetic reconstruction of mammalian MHC class I genes with the CrHV2 homolog. The CrHV2 branch length was truncated to fit in the figure, and the total branch length is indicated in parentheses. The scale bar represents 0.1 substitutions per site. The numbers at nodes represent posterior probabilities, and values below 80% are not shown. This tree shows that the CrHV2 homolog and the genes of *M. musculus* and *R. norvegicus* group together with a 100% probability. (B) Graph depicting a collation of a number of selection analysis results. The positions of known domains are labeled above the graph. The gray line indicates the *dN*/*dS* value obtained for each residue. Sites shown to be under significant positive selection according to the Bayes empirical Bayes (BEB) analysis (a method to calculate the posterior probability that a site belongs to a *dN*/*dS* value category, using a 95% threshold) are indicated by gray dots along the line. The green (host) and blue (virus) lines represent the results of the sliding-window branch test analysis. The thick lines indicate areas where the branch test rejected the null hypothesis of a single *dN*/*dS* value according to a likelihood ratio test. (C) Structural model of the MHC homolog (gold) with the best-matching template (mouse MHC) in the database (blue). There is a helix at both termini that could not be modeled, suggesting that it could have been lost in the viral copy (the corresponding region is not homologous). This suggests that the viral copy does not function as a transmembrane protein, since the C-terminal helix is needed for this. (D) The same structural model as in panel C, but showing only the antigen binding cleft, colored according to the *dN*/*dS* value for each residue. Nearly all of the positively evolving sites are found in this region.

(*Callithrix jacchus*), which speciated after the divergence of their common ancestor from spider monkeys (*Ateles* sp.). This may mean that the gene was acquired by the SaHV2 ancestor after the squirrel monkey-spider monkey split but before the squirrel monkey-marmoset split, ~20 million years ago (mya) (16). This result suggests that the capture event may have occurred in a small, 3-million-year window, when New World monkeys were undergoing a major radiation that separated the major clades (Atelidae, Cebidae, and Pithicidae) (17).

The genomes of HHV8 and *Macacine gammaherpesvirus 5* (MaHV5) contain homologs of the IL-6 gene, which encodes a secreted cytokine known to have a variety of functions, ranging from roles in immunomodulation to oncogenesis (18). The phylogenetic analysis of these genes revealed that the sequences are embedded within a clade that includes the mouse, rat, and tree shrew (Fig. 4A). This is surprising given that both HHV8 and MaHV5 are primate-infecting viruses, which suggests that the gene may have been acquired from a nonprimate host, consistent with a history of cross-species transmission of rhadinoviruses (19).

CD59 glycoprotein and core 2 β-1,6-*N*-acetylglucosaminyltransferase–mucin (C2GnT-M) genes. SaHV2 encodes a homolog of the mammalian cell surface glycoprotein CD59, first identified during the genome sequencing project of the virus in 1992. The results of our phylogenetic analysis demonstrate that the gene was acquired through HGT, since the viral homolog resolves next to *Saimiri boliviensis* with a high posterior probability, to the exclusion of *Callithrix jacchus* CD59 (Fig. 5A). We can therefore conclude with confidence that the gene was captured after the *Callithrix*-

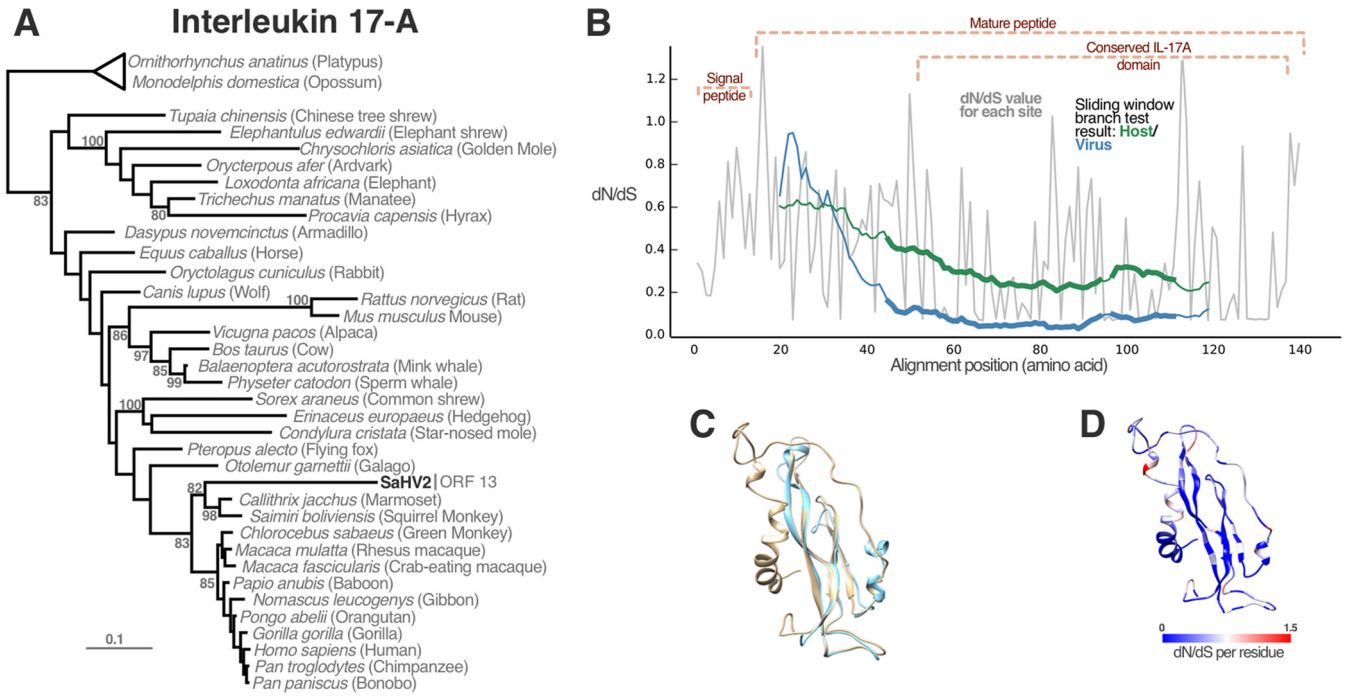

FIG 3 Host-derived IL-17 genes in MaHV5, HHV8, and SaHV2. (A) Bayesian phylogenetic reconstruction of mammalian IL-17 genes and a homologous sequence from SaHV2. The scale bar represents 0.1 substitutions per site. The numbers at nodes represent posterior probabilities, and values below 80% are not shown. The tree shows that the viral gene groups most closely to the genes from the New World monkeys *C. jacchus* and *S. boliviensis*, the latter being SaHV2's current host. (B) Graph depicting a collation of a number of selection analysis results. The gray line indicates the *dN/dS* value obtained for each residue. No sites under significant positive selection according to the BEB analysis were identified. The green (host) and blue (virus) lines represent the results of the sliding-window branch test analysis. Thick lines indicate areas where the branch test rejected the null hypothesis of a single *dN/dS* value according to a likelihood ratio test. The conserved domains and other regions in IL-17 are annotated in brown. (C) Structural model of the IL-17 homolog (gold) with the best-matching template (human IL-17) in the database (blue), showing precise overlap for most of the structure. (D) The same structural model as in panel C, but without the template model and colored according to the *dN/dS* value for each residue.

Saimiri divergence (~20 mya) (16), from either the current host (*Saimiri*) or a closely related monkey that may be susceptible to infection. This is also supported by the fact that the gene is absent in AtHV3, the virus most closely related to SaHV2.

The *Bovine gammaherpesvirus 4* (BoHV4) Bo17 gene was previously identified as a host-derived C2GnT-M gene that is involved in glycan synthesis. It has already been established that the gene is likely derived from an ancestor of the African buffalo (*Syncerus caffer*) (20), which has since been shown to be the most likely natural host of BoHV4, which was subsequently transmitted to cattle multiple times (21). The tree reconstructed in our analysis is consistent with these previous findings, placing the BoHV4 C2GnT-M gene next to the African buffalo gene with a high posterior probability (Fig. 5B).

Dihydrofolate reductase (DHFR), CD200 immunoglobulin, and CCL3 genes. Several of the host-derived genes investigated were found in multiple viruses, suggesting that they were acquired before the viruses diverged. The genomes of MaHV5 and HHV8 include ORFs with similarity to the small cellular chemokine C-C motif ligand 3 (CCL3), which plays a role in the inflammatory response. Three such genes were identified in HHV8, and phylogenetic reconstruction places the clade of viral genes within primates, with over 80% probability, but their exact position is not well supported (Fig. 5C) (posterior probability values are between 0 and 1 but are expressed as a percentage throughout the paper). The fact that the HHV8 genes group together, to the exclusion of the MaHV5 copy, with a 97% probability, indicates that they were duplicated within the herpesvirus after capture.

Three rhadinoviruses contain a homolog of DHFR, which is a universal cellular enzyme that catalyzes the conversion of dihydrofolate to tetrahydrofolate (22). Tetrahydrofolate is required for purine and thymidylic acid synthesis, as well as synthesis of

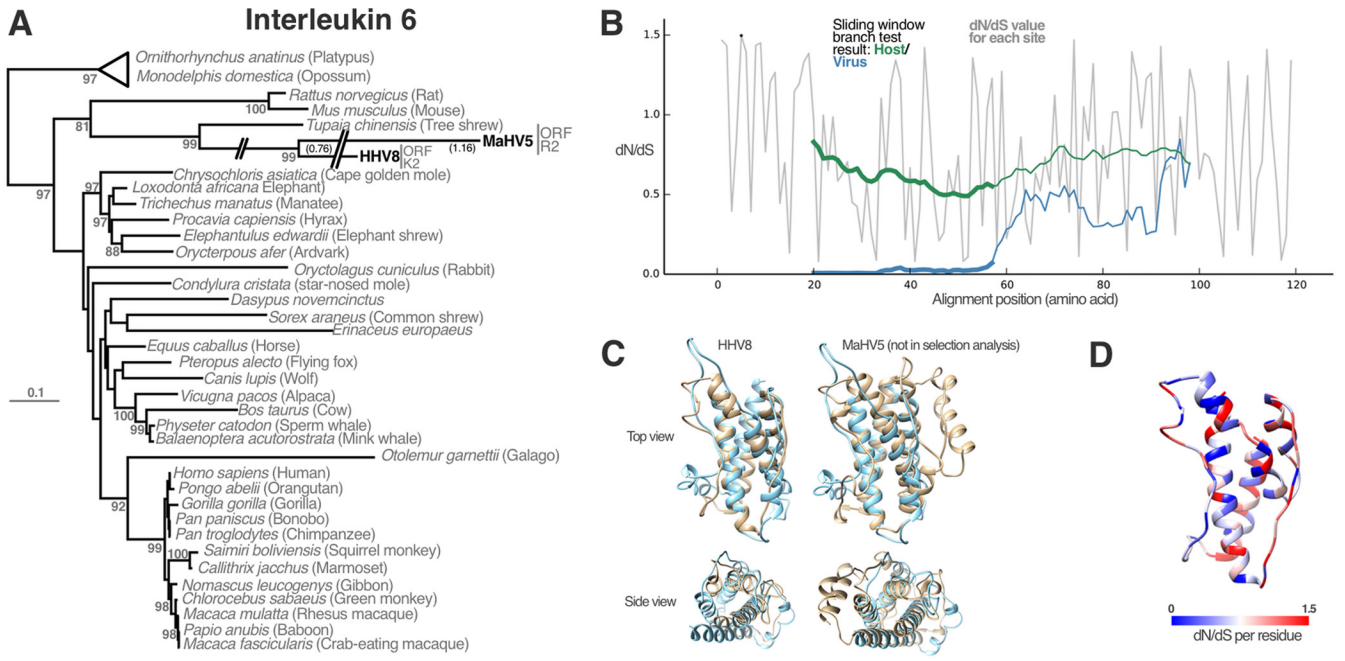

FIG 4 Host-derived IL-6 genes in MaHV5, HHV8, and SaHV2. (A) Bayesian phylogenetic reconstruction of mammalian IL-6 genes and homologous sequences from HHV8 and MaHV5. The scale bar represents 0.1 substitutions per site. The numbers at nodes represent posterior probabilities, and values below 80% are not shown. The tree shows that the viral genes group most closely to the tree shrew gene (*Tupaia chinensis*), with 99% probability, despite belonging to viruses that infect primates. This may indicate either that a cross-species transfer occurred in the history of the virus or that long-branch attraction resulted in this unusual topology. (B) Graph depicting a collation of a number of selection analysis results, not including the MaHV5 gene, since it could not be aligned for the entire length (219 gaps in a 384-nucleotide alignment). The gray line indicates the *dN*/*dS* value obtained for each residue. Sites shown to be under significant positive selection according to the BEB analysis are indicated by gray dots along the line. The green (host) and blue (virus) lines represent the results of the sliding-window branch test analysis. Thick lines indicate areas where the branch test rejected the null hypothesis of a single *dN*/*dS* value according to a likelihood ratio test. (C) Structural models of both IL-6 homologs (gold) with the best-matching template (mouse IL-6) in the database (blue). Note that as well as sequence divergence, the MaHV5 structure is also more divergent than the HHV8 counterpart. (D) Structural model of HHV8, as in panel C, but without the template model and colored according to the *dN*/*dS* value for each residue. Residues with a *dN*/*dS* value above 1 are distributed throughout the structure, with no discernible pattern.

amino acids such as glycine and cysteine (22). Although the gene is situated in variable region A for MaHV5 and SaHV2, the HHV8 DHFR homolog is located downstream of this, in variable region B. If all three genes originated from the same capture event, then the DHFR gene in HHV8 must have either been moved through recombination between variable regions A and B or been duplicated and then lost from region A. However, there is a 23-amino-acid carboxyl-terminal region in HHV8 DHFR that is absent in MaHV5 and SaHV2, a difference that has been pointed out as supportive of a separate capture scenario to explain the distinct locus (23). Although the tree in this study shows that the genes cluster together with a 98% posterior probability, it is not possible to rule out a separate capture, since intermediate lineages may not have been sampled (Fig. 6A).

The two viruses also carry a gene with similarity to the CD200 gene, encoding a membrane glycoprotein of the immunoglobulin superfamily also known as OX-2. Experimental evidence has shown that purified HHV8 OX-2 stimulates the production of inflammatory cytokines in myeloid-lineage cells (24). Phylogenetic reconstruction demonstrates conclusively that the gene is primate derived, with a high posterior probability (96%) (Fig. 7A). The topology indicates that the viral gene is most similar to New World monkey host genes (from *Saimiri boliviensis* and *Callithrix jacchus*; 92% posterior probability), although the extremely long branches may have influenced their placement.

FGAMS and complement control protein (CCP) genes. Although this study is not focused on genes that were likely captured before the emergence of rhadinoviruses, we examined the evolution of two such genes due to their extensive postcapture duplication and rearrangement in rhadinoviruses. Phosphoribosylformylglycinamidine syn-

FIG 5 Phylogenetics and structural models of captured genes in HHV8, BoHV4, and SaHV2. This figure contains the genes from our main analysis for which a sliding-window analysis was not conducted, because of a short alignment length in the case of CD59 and CCL3 and because the overall branch test was not significant for C2GnT-M. (A) Bayesian phylogenetic reconstruction of mammalian CD59 glycoprotein genes and homologous sequences from SaHV2. The scale bar represents 0.1 substitutions per site. The numbers at nodes represent posterior probabilities, and values below 80% are not shown. The tree shows that the viral gene groups with the *S. boliviensis* gene, with 97% probability, indicating that the gene is most likely derived from this monkey, which is the virus's current host. The relatively short branch length is also an indication that the capture occurred recently. (B) Bayesian phylogenetic reconstruction of mammalian C2GnT-M genes with a homolog from BoHV4. The scale bar represents 0.1 substitutions per site. The numbers at nodes represent posterior probabilities, and values below 80% are not shown. This tree confirms previous studies that have reconstructed trees with comparable topologies. This has previously been interpreted as evidence that the origin of the gene is *S. caffer*. (C) Bayesian phylogenetic reconstruction of mammalian CCL3 genes, homologous genes from MaHV5, and three copies of the gene in HHV8. The scale bar represents 0.1 substitutions per site. The numbers at nodes represent posterior probabilities, and values below 80% are not shown. (D to F) Structural models of CD59, C2GnT-M, and CCL3 are represented in gold. The best-matching database templates for CD59 and C2GnT-M were human and mouse C2GnT-L, respectively. In the case of the CCL3 homologs, the best-matching template was the human thymus and activation-regulated chemokine (TARC).

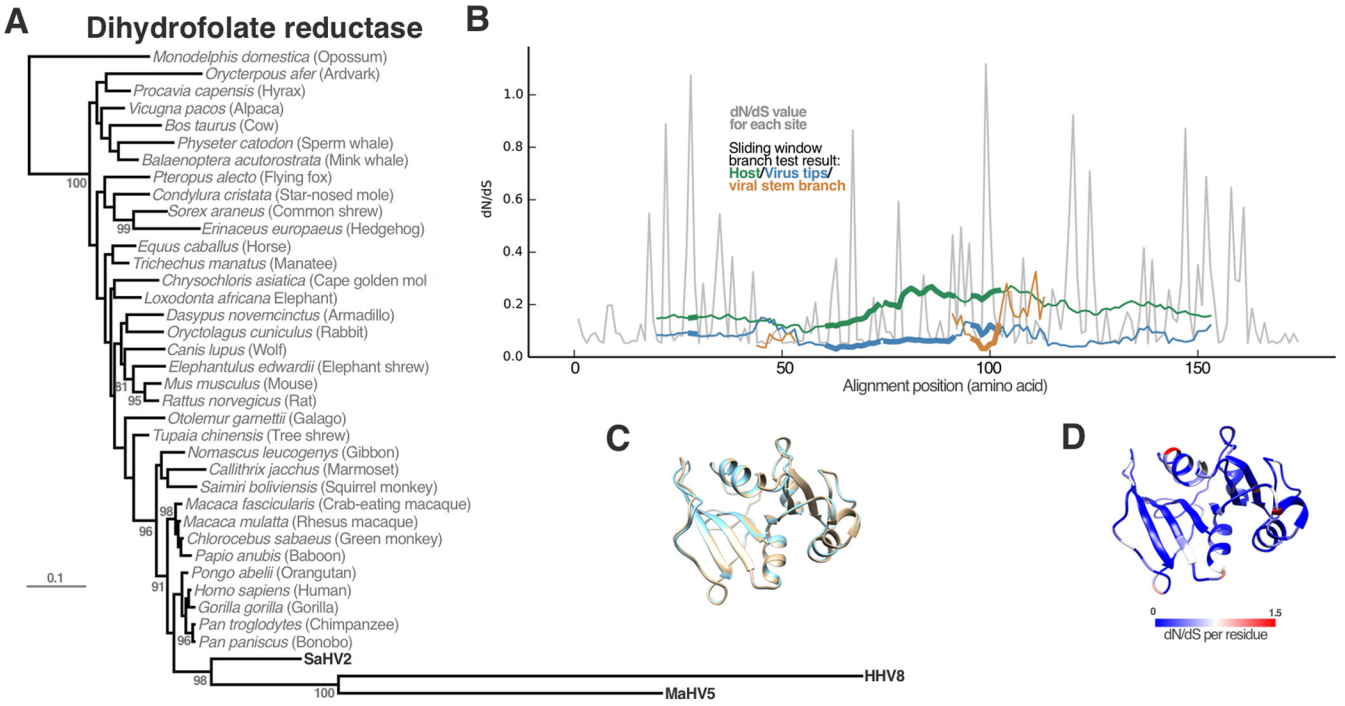

FIG 6 Host-derived DHFR genes in SaHV2, HHV8, and MaHV5. (A) Bayesian phylogenetic reconstruction of mammalian DHFR genes and homologous sequences from SaHV2, HHV8, and MaHV5. The scale bar represents 0.1 substitutions per site. The numbers at nodes represent posterior probabilities, and values below 80% are not shown. The tree shows that the viral genes group with primate genes, with 91% probability, and are probably most closely related to great ape genes, as there is a posterior probability of 96% that the virus clade and the clade for great apes are sister clades. (B) Graph depicting a collation of a number of selection analysis results. The gray line indicates the *dN*/*dS* value obtained for each residue. No sites under significant positive selection according to the BEB analysis were identified. The green (host), blue (virus), and orange (viral stem branch) lines represent the results of the sliding-window branch test analysis. Thick lines indicate areas where the branch test rejected the null hypothesis of a single *dN*/*dS* value according to a likelihood ratio test. *dN*/*dS* estimates of infinity for the viral stem branch are not shown, for clarity. (C) Structural model of the DHFR homolog (gold) with the best-matching template (human dihydrofolate reductase) in the database (blue), showing precise overlap for the entire structure. (D) The same structural model as in panel C, but without the template and colored according to the *dN*/*dS* value for each residue.

thase (FGAMS) is a purine synthesis enzyme which was captured early in the evolution of gammaherpesviruses, as it is shared by the whole group. Among rhadinoviruses, however, a number of duplications, deletions, and translocations have occurred, though they have not been investigated fully. We conducted a phylogenetic reconstruction of available gammaherpesvirus FGAMS genes to trace the series of genomic changes that led to the current state of *Rhadinovirus* FGAMS (Fig. 8A and B). Consistent with previous observations, the tree shows that after the ancestral capture, two initial duplication events occurred, resulting in three copies in the ancestor of macaviruses, percaviruses, and rhadinoviruses. Both new copies appear to have been lost in the MaHV5-HHV8 ancestor but were duplicated a second time in the ancestor of CrHV2, *Murid gammaherpesvirus 4* (MuHV4), and *Wood mouse herpesvirus* (WMHV). These rodent viruses also appear to have lost the ancestral FGAMS gene, since their copies are more closely related to one of the copies that arose in the ancient double duplication. This suggests that a translocation event may have occurred at the time of this secondary event. This second duplicate was itself duplicated in the MuHV4-WMHV ancestor, since both genomes contain a third copy that is most closely related to it.

A homolog of the host gene that encodes CCP is found in all rhadinoviruses except BoHV4. As with several members of the complement system (a subset of innate immunity), CCP contains modular sushi repeats, which are conserved, ~60-amino-acid domains that exist in a range of complement and adhesion proteins (25). In rhadinoviruses, the number of sushi domains is variable in the different homologs (ranging from 4 to 8). The rhadinoviral homologs are most similar at the sequence level to the conserved mammalian complement component 4 binding protein (determined by BLAST searching) but could not be aligned with confidence due to low sequence

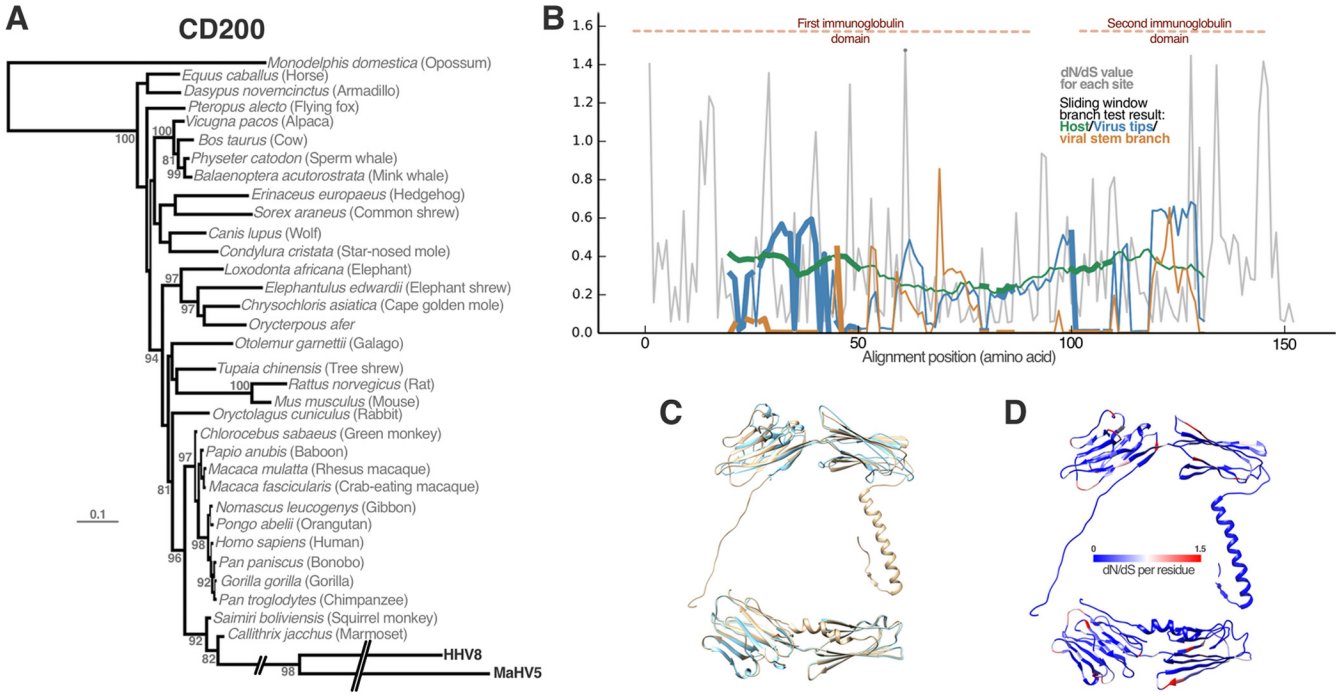

FIG 7 Host-derived CD200 genes in SaHV2, HHV8, and MaHV5. (A) Bayesian phylogenetic reconstruction of mammalian CD200 genes and homologous sequences from HHV8 and MaHV5. The scale bar represents 0.1 substitutions per site. The numbers at nodes represent posterior probabilities, and values below 80% are not shown. The viral genes are most closely related to New World monkey host genes from *C. jacchus* and *S. boliviensis*, with a 92% probability, and are more closely related to the *C. jacchus* gene, with an 82% probability. (B) Graph depicting a collation of a number of selection analysis results. The gray line indicates the *dN*/*dS* value obtained for each residue. One site under significant positive selection according to the BEB analysis is indicated by a gray dot. The green (host), blue (virus), and orange (viral stem branch) lines represent the results of the sliding-window branch test analysis. Thick lines indicate areas where the branch test rejected the null hypothesis of a single *dN*/*dS* value according to a likelihood ratio test. *dN*/*dS* estimates of infinity for the viral stem branch are not shown, for clarity. (C) Structural model of HHV8 (upper) and MaHV5 (lower) homologs (gold). The best-matching template in the database was mouse CD200 (blue). (D) Structural models of HHV8 and MaHV5 as in panel C, but without the template model and colored according to the *dN*/*dS* value for each residue.

identity. Furthermore, it is reasonable to assume that the CCP genes are orthologs, because they are all found at roughly the same locus and are partially syntenic (Fig. 1). HHV8, MuHV4, and SaHV2 each have one gene encoding a protein with similarity to CCP containing 4 sushi domains, but none of the viruses most closely related to each of these viruses possess a CCP ortholog with the same number of sushi domains. Based on a high posterior probability grouping the genes in the phylogeny, we can conclude that they originated from a single capture of a gene that likely encoded 4 sushi domains, based on the domain organization of a similar CCP host gene (CD46). Based on these conclusions, AtHV3 appears to have lost a single sushi domain since it diverged from SaHV2. In MaHV5, an apparent duplication resulted in a gene that is twice as long as the HHV8 copy (encoding 8 sushi domains). A similar duplication seems to have occurred in CrHV2 compared to MuHV4, but in this case there are two separate ORFs separated by several genes, with each missing the sequences for two sushi domains.

To investigate the possible evolutionary history of duplication and loss of CCP genes, we conducted a cophylogeny analysis using JANE 4.0 (Fig. 8D and E). This allowed us to reconstruct the possible series of events based on the respective phylogenetic relationships among the genes and the topology of the gene tree relative to that of the virus tree. This revealed one most parsimonious solution when HGT between the viruses is permitted and one solution where no HGT is allowed. In both instances, we used the default cost settings for duplication, loss, and failure to diverge. Under the latter scenario, only SaHV2 and AtHV3 have retained the ancestrally captured gene, which was replaced with subsequent duplicates in all other viruses. The first duplication gave rise to the 5′ end of the MaHV4 CCP gene as well as the copy in CrHV2

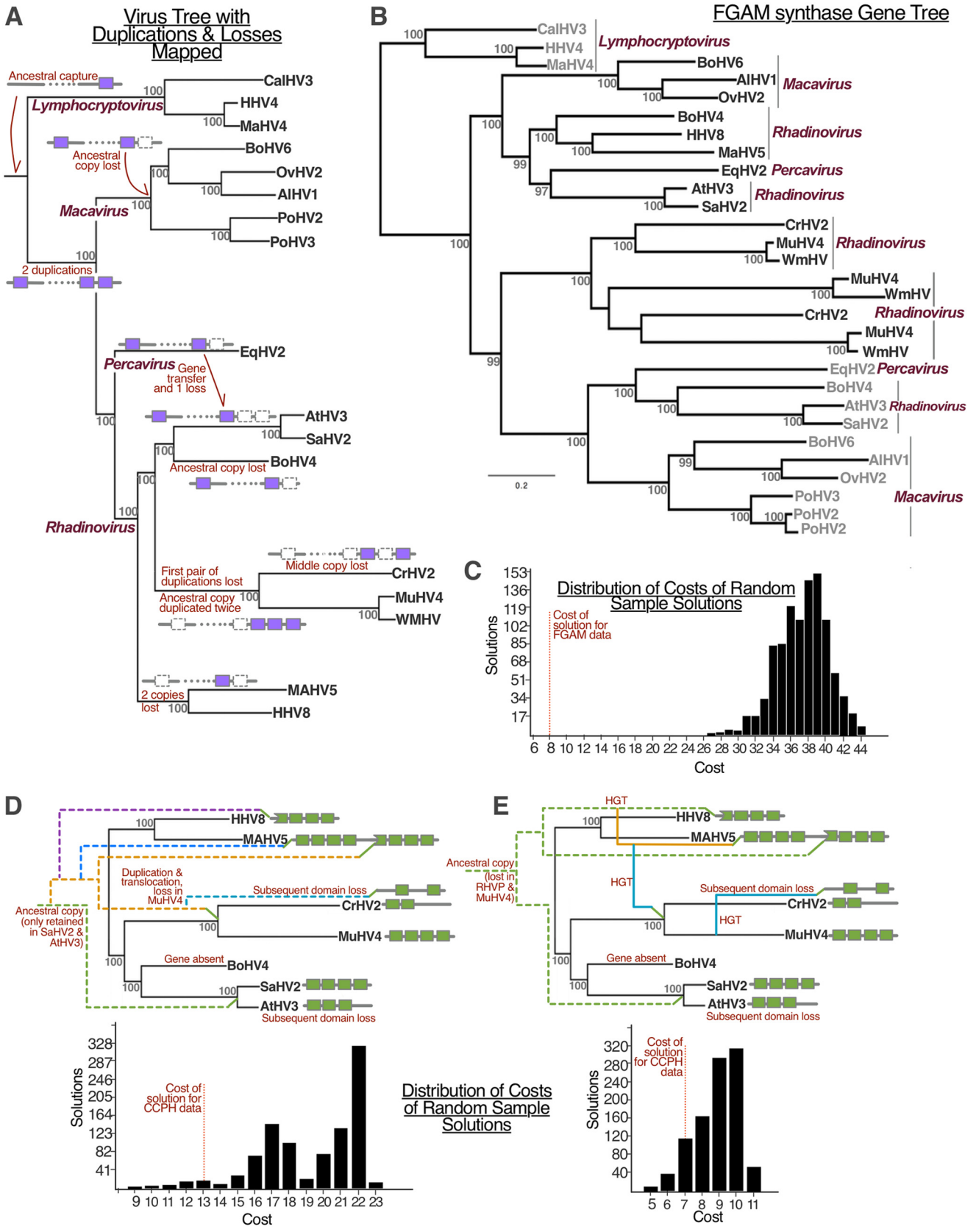

FIG 8 Cophylogeny analysis of FGAM synthase and CCP genes. (A) Results of the cophylogeny analysis of FGAMS homologs in rhadinoviruses. The tree shown is a maximum likelihood phylogeny reconstructed from a concatenation of the six core herpesvirus genes, with reconstructed duplication, rearrangement, and horizontal transfer events superimposed as genome sketches. Although the viral sequences are too divergent to align with host genes, a series of

(Continued on next page)

and MuHV4. This copy underwent a further two rounds of duplication, leading to the HHV8 copy and the 3′ end of the MaHV5 gene. In CrHV2, a lineage-specific duplication gave rise to the second ORF, after which both genes lost the sequences for two sushi domains. Alternatively, under a model where HGT is permitted, the analysis reveals that the ancestral copy gave rise to the contemporary gene in SaHV2, AtHV3, and HHV8 and to the 5′ end of the MaHV5 gene. The 3′ end of the MaHV5 CCP gene was the result of a duplication and transfer of a copy from HHV8, which itself also transferred to the MuHV4-CrHV2 ancestor. Thereafter, the extra copy in CrHV2 originated from a similar duplication and transfer from MuHV4.

Other genes. Some noncore genes exhibited very low similarities to cellular genes, and as such could not be used as part of a phylogenetic analysis to determine the host group of origin with statistical confidence. For example, all of the rhadinoviruses, apart from MuHV4 and CrHV2, carry a CASP8-like gene, which in hosts encodes a protease involved in programmed cell death. Although the phylogeny was inconclusive with regard to the particular host lineage of origin, it revealed that the gene does group within mammalian homologs, with 93% probability (see Fig. S1A in the supplemental material). Similarly, HHV8 and MaHV5 share a number of interferon-like genes that were probably acquired after their divergence from other rhadinoviruses. Interestingly, these genes are located within a single stretch amid core block 4 that is otherwise syntenic in nearly all herpesviruses (Fig. 1). There are also genes that could be aligned to host homologs, but phylogenetic analysis was inconclusive. For example, the BoHV4 genome includes a small hypothetical open reading frame with similarity to an acyl-coenzyme A (acyl-CoA) thioesterase gene, which is part of a family of cellular genes that encode enzymes involved in metabolism (Fig. S1B).

HHV8, MaHV5, SaHV2, and AtHV3 all possess another nucleotide metabolism gene, namely, a thymidylate synthase (TS) gene. It is in the same location in MaHV5 and HHV8 but situated near the right terminus of the genome in SaHV2 and AtHV3 (Fig. 1). Although the phylogeny exhibited poor support for most nodes, the tree reveals that the TS genes in SaHV2 and AtHV3 have much shorter branches than those for the HHV8 and MaHV5 genes, suggesting that they originated from a distinct and more recent capture (Fig. S1C). The fact that the pairs of TS genes are at opposite locations in their respective genomes also supports this hypothesis.

The leftmost HHV8 gene, K1, encodes a transformation-associated glycoprotein that performs roles in signal transduction, possibly involved in immunomodulation and perpetuating cell survival (26). Based on a BLAST search against herpesviruses, K1 does not appear to have homologs in other genomes, except for very low sequence similarity to MaHV5 R1 (28% amino acid identity, with ~50% query coverage). While we cannot identify host sequences with high similarity to K1, R1 of MaHV5 is most similar to the N terminus of the low-affinity receptor for the Fc fragment of immunoglobulin G in primates (31% amino acid identity to *Callithrix jacchus*, with 41% coverage). It is possible that this gene is the original host homolog of the viral ORF, consistent with functional studies of K1, which is similar to R1 (albeit distantly). It has been shown that K1 can cause downregulation of the B cell antigen receptor complex (BCR), which is

FIG 8 Legend (Continued)
duplications, losses, and rearrangements occurred in the clade. The viral tree is shown with reconstructed events superimposed at each node. The sketches represent a simplified viral genome with FGAMS genes at either end (purple boxes). Dotted lines indicate a loss event, and the duplications can be deduced based on a comparison of a node with the previous node. (B) Phylogenetic tree of all FGAMS genes from rhadinoviruses (including copies). Dark taxon labels indicate 3′-end loci. The scale bar represents 0.2 substitutions per site. (C) Histogram of the costs of random sample solutions, showing the cophylogeny reconstruction for FGAMS at a cost of 8, compared to the next best random cost of 27 ($P < 0.0001$). We implemented a cost scheme of 0, 1, 2, 1, and 1 for cospeciation, duplication, loss, host switching, and failure to diverge, respectively. (D and E) Results of the cophylogeny analysis of CCP homologs, superimposed on the virus tree as a cladogram (the branch lengths of colored dotted lines are arbitrary). Each of the sushi domains within CCP viral homologs (green boxes) is represented as a cartoon next to each taxon label. The best (only) solutions are shown with (E) and without (D) horizontal gene transfer allowed. The costs for the CCP solutions with and without HGT allowed lie within the distribution of costs for random samples ($P = 0.151$ and $P = 0.029$, respectively). The same cost scheme as that in panel C was used. Virus abbreviations are as follows: CalHV3, *Callitrichine herpesvirus 3*; HHV4, *Human herpesvirus 4* (also known as *Epstein-Barr virus* [EBV]); HHV8, *Human herpesvirus 8*; MaHV5, *Macacine gammaherpesvirus 5*; BoHV4/6, *Bovine gammaherpesvirus 4/6*; SaHV2, *Saimiriine gammaherpesvirus 2*; AtHV3, *Ateline gammaherpesvirus 3*; MuHV4, *Murid gammaherpesvirus 4*; CrHV2, *Cricetid gammaherpesvirus 2*; EqHV2, *Equine herpesvirus 2*; WMHV, *Wood mouse herpesvirus*; and PoHV2/3, *Porcine herpesvirus 2/3*.

known to occur via interaction of the N terminus with the μ chain of BCR (27). This would explain why sequence conservation has been maintained at the N terminus.

We were unable to robustly estimate a phylogeny of the R1 gene with host homologs, probably due to the effects of long-branch attraction (where a large number of mutations can increase the chance of spurious branch placement). Nonetheless, the MaHV5 gene groups with placental mammal genes, with a high posterior probability (Fig. S1D). K1 and R1 are situated in the same genomic position, and both contain immunoglobulin-like domains. Their synteny may be interpreted as evidence that the genes are highly divergent orthologs that were acquired from the ancestral primate host, but we cannot rule out the possibility that they are independent acquisitions. The latter possibility is supported by their extreme divergence from one another compared to the rest of the genomes (which are otherwise easily aligned). Even among HHV8 K1 genes, the sequence diversity is extremely high among different strains, suggesting that there is a mechanism of diversifying selection (28), and the same evolutionary pressure can be invoked to explain the repeated capture of the same gene. An interesting possibility in the case of K1/R1 is that the variability may have been selected in order to mimic or avoid a particular HLA/MHC repertoire (28).

Postcapture evolutionary changes. (i) Analysis of selection. We determined that the selection pressure imposed on viral genes was significantly different from that on their cellular homologs (likelihood ratio test; $P \leq 0.05$) for seven of the eight genes tested (C2GnT-M was the exception). Using the branch model implemented in CODEML, we found that the viral genes exhibited lower ratios of nonsynonymous to synonymous evolutionary changes (*dN*/*dS*) than those of their cellular counterparts (Table S1). The *dN*/*dS* ratio reflects the balance between neutral, deleterious, and beneficial mutations. Values significantly lower than 1 indicate the presence of purifying selection, whereas *dN*/*dS* values above 1 indicate that beneficial changes are likely being maintained (i.e., positive selection). For genes for which multiple viruses contained a homolog (DHFR, CD200, and CCL3), we found that the drop in *dN*/*dS* value was evident whether the model included separate estimates for the viral stem and crown group branches or included both as a single estimate.

We also identified 1 to 22 sites under positive selection in 6 of the 8 genes tested (Table S1). Since the test of positive selection considers the whole alignment for all taxa, the results primarily reflect the evolutionary history of the host genes, since they represent the majority of sequences in the data. This indicates that the genes are evolving adaptively in host genomes and that the values are likely to be an underestimate of the true number of positively selected sites due to the limitations of selection analyses (see Materials and Methods). We therefore tested whether the overall *dN*/*dS* reduction in viral branches according to the branch model could be due to an artifact of these positively evolving sites in the host that was biasing the *dN*/*dS* estimates. We implemented a modification to the branch model whereby these sites were excluded, and we found that the drop in *dN*/*dS* was still significant, albeit with a slightly lower magnitude (Table S2).

As well as considering the possible biasing effect of sites evolving under positive selection, we wanted to explore how different regions of a gene were contributing to the findings of overall *dN*/*dS* reduction. We therefore implemented a sliding-window modification of the branch model, which allowed us to consider whether subsections of the gene would return a significant reduction in *dN*/*dS*. In this analysis, the *P* value was used as a proxy for the magnitude of the effect and therefore did not require correction for multiple testing. Interestingly, we found that the difference in *dN*/*dS* values between cellular and viral counterparts is nonuniform along the lengths of the genes. In the case of MHC genes, the test revealed that the viral homolog *dN*/*dS* value is significantly lower only for the regions that correspond to the alpha-1 and alpha-2 domains (Fig. 2B). Together these form the antigen binding cleft, which is where the majority of positively selected sites were detected. Moreover, the viral *dN*/*dS* values

within the alpha-1/2 region are lower than the values of the alpha-3 region, for which the virus and cell *dN*/*dS* values cannot be statistically distinguished (Fig. 2B).

For IL-17A genes, we found that the significant difference in *dN*/*dS* values is localized to the conserved IL-17 domain and signal peptide, and for both IL-6 and DHFR genes, the drop in *dN*/*dS* value is limited to roughly half of the gene (Fig. 4B and 6B, respectively). Since a homolog of DHFR is found in more than one rhadinovirus, the sliding-window branch test also revealed a small stretch of significantly different *dN*/*dS* values for host branches, viral tips, and the viral stem branch (Fig. 6B). For the CD200 homolog, we found several regions across the gene that exhibit a significant difference in *dN*/*dS* values between host and viral branches. Similar to the DHFR results, the 3-model branch test also revealed multiple regions with a different *dN*/*dS* value for the viral stem branch (Fig. 7B). Unlike the analyses of the rest of the genes, however, the CD200 analysis showed that the differences in *dN*/*dS* values are not contiguous but rather interspersed between regions where a significant difference could not be detected between virus and host values. Furthermore, the sliding-window test also revealed that the genes are evolving with comparatively higher *dN*/*dS* values for the virus, with peaks at three different regions (Fig. 7B).

(ii) Sequence composition and structural modeling. We further investigated the differences between viral genes and their host homologs by comparing nucleotide frequencies (mono- and dinucleotide), CpG bias, and gene length in a principal component analysis. As was the case for the branch site analysis, the results showed that there was a substantial difference in composition for all the genes except the C2GnT-M homolog (Fig. 9). We also included all the genes in each viral genome in the analysis. This revealed that the sequence composition of the captured viral genes was moving away from that of the hosts from which they were derived and toward the characteristics of the virus genome (Fig. 9). In contrast, structural modeling of the most conserved regions of viral homologs showed that the sequences are compatible with the structures of their host homologs (Fig. 2 to 6). Furthermore, by mapping the selected sites onto the modeled structures, we were able to corroborate their functional relevance and to minimize the possibility that the results of the selection analysis are random false-positive results. This is particularly striking in the case of the MHC homolog in CrHV2, where the majority of selected sites are localized to the antigen binding cleft (Fig. 2D).

DISCUSSION

We investigated the history and evolution of host-derived genes in rhadinoviruses and identified a common evolutionary strategy in captured genes. We have shown that all the host-derived viral genes investigated in this study are under relatively stronger purifying selection than that on their host homologs, despite the strong transformative forces imposed on them. This includes the fact that, on average, genes are evolving 1 to 2 orders of magnitude faster (in terms of substitutions per nucleotide per year) in viruses than in their hosts (29, 30).

This high evolutionary rate is reflected in the long branches of rhadinoviral homologs in phylogenies. Moreover, selective forces have driven the base composition away from that of the hosts (Fig. 9). We were nonetheless able to identify the host lineage of origin for eight rhadinoviral genes. For example, the trees in Fig. 6 and 7 both indicate that the cellular homologs originated from primates, and more specifically, the CD59 homolog in SaHV2 derived from a *Saimiri* species, with a posterior probability of 97% (Fig. 5A).

Together these findings suggest that the conservation of the acquired genes is being selected for via a significant relative drop in *dN*/*dS* value compared to the host homologs. The relative drop in *dN*/*dS* value might be interpreted as an artifact of the viruses' higher evolutionary rate resulting in an increase in synonymous changes that would artificially decrease the overall ratio. However, there is no evidence to suggest that the higher mutation rate would influence nonsynonymous mutations differently.

Conservation of function via higher selective constraints is supported by our structural modeling analyses revealing that conserved domains are maintained (i.e., struc-

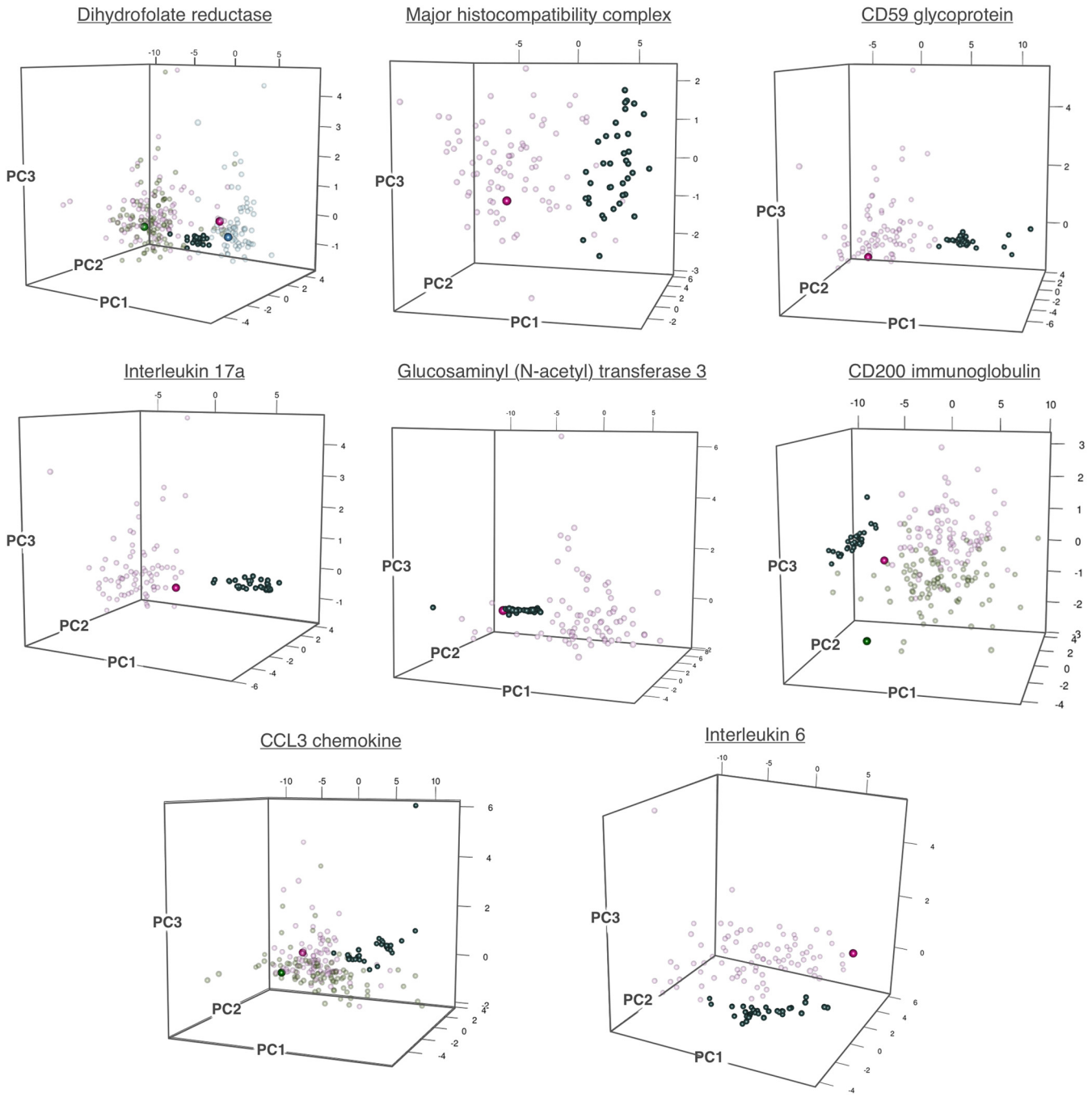

FIG 9 Principal component analysis of sequence composition, length, and CpG bias. The results of principal component analysis of the eight main genes investigated in this study are presented. The variables used were all 4 mononucleotide and 16 dinucleotide frequencies, in addition to CpG bias and gene length. For each gene, the scores for the first three principal components (accounting for most of the variance) were plotted for the mammalian host genes as small, solid, dark-green points. All the data for the viral genes in the genome are shown in either pink, blue, or lime green, with the host-derived gene in question highlighted as a larger, dark point of the same color. In the case of the dihydrofolate reductase, CD200, and CCL3 genes, host-derived homologs were observed in multiple rhadinoviruses, so each virus was assigned a different color. For all of the genes except for C2GnT-M, there is a clear distinction in variable space between the viral homologs and the host genes from which they originated. See Table S4 in the supplemental material for a detailed breakdown of the variable loadings for each principal component.

tural similarity is not simply due to recent acquisition). While such modeling approaches are not always accurate predictions of the true protein structure, this result demonstrates that the selective forces have not significantly influenced secondary or tertiary structure. Furthermore, by mapping the changes in *dN*/*dS* value to the structural models, we can reveal how the patterns of selection pressure correspond to the

different regions of the protein in three-dimensional (3D) space (rather than only considering their location along the sequence).

Gene capture as an evolutionary shortcut. Hosts also capture viral genes as part of the arms race, with EVEs repurposed as antiviral immune genes or regulators thereof (6, 10, 31). Both host-derived viral genes and virally derived host genes are an "evolutionary shortcut," circumventing the need for incremental adaptation (10). This is consistent with finding captured genes primarily in dsDNA viruses: in addition to being able to accommodate more genes (due to size), they have the lowest viral mutation rates (32) and therefore the smallest rate advantage, although there is evidence that host demographics and selective pressure can markedly increase dsDNA virus mutation rates (33). While this evolutionary shortcut strategy is rare in hosts, the large number of lineage-specific cellular genes in viruses indicates a highly recurring event that occurs through various mechanisms (34). For example, poxviruses—also large dsDNA viruses—employ "genomic accordions," where genomic regions expand transiently through gene duplication as a crude mechanism of enhancing expression levels (35). Once the selection pressure from the host diminishes, the virus can contract by losing these costly extra genes, but it can retain those that may have evolved novel adaptive phenotypes (35).

Purifying selection in captured genes as one weapon in the evolutionary arms race. The selection tests we developed demonstrate that the rhadinoviral strategy of preservation of function is achieved through a relative increase in purifying selection (Fig. 2 to 4 and 6), despite the fact that host immune genes evolve at higher rates than those of other host genes (36, 37). A possible interpretation of our results is that the fast evolutionary tempo in hosts is an adaptive response to multiple viral threats. However, in the rhadinoviral context, a captured host gene has a much narrower target to evade—presumably the host species from which it came, or a very closely related species. The selection shift can be seen as the underlying mechanism behind mimicry, the process by which a pathogen uses a molecular mimic to imitate and thereby subvert host processes to benefit itself (38).

We dissected the dynamics of this selection shift by developing a sliding-window approach to the CODEML branch model. Many sliding-window *dN*/*dS* techniques are susceptible to false-positive results and problematic due to a lack of correction for multiple testing (39). However, having established an overall trend by using the standard branch test, we used a sliding window that implements the likelihood ratio test only as a *post hoc* exploratory tool. This allowed us to evaluate the heterogeneity of the trend at subsections of the gene where the significance of the *P* value correlates with the magnitude of the effect. For example, for the IL-6 gene, the largest difference in *dN*/*dS* value is localized to the first half of the sequence (Fig. 4). This may help to pinpoint the specific functional effects that drove the acquisition of the gene, since IL-6 has been implicated in a range of different processes that may be useful to a virus (18).

The adaptive nature of this selection shift is clearest for the IL-17 and MHC homologs (Fig. 1 and 3, respectively), for which the drop of *dN*/*dS* value is specific to the known functional domains. For the MHC homolog, these domains are also where most of the 22 positively selected sites were detected, which likely represents the positive selection undergone by host genes that are overrepresented in the alignment (Fig. 1; see Table S1 in the supplemental material). Vertebrate MHC genes are among the most rapidly evolving vertebrate genes, and there are multiple hypotheses that together explain their paradoxical levels of genetic diversity (40, 41). MHC molecules function by presenting bound antigens to T cells to elicit an immune response, and the high diversity of MHC molecules is driven by the evolutionary pressure to recognize a wide range of antigens (41). A counterstrategy employed by viruses is the downregulation of host MHC molecules to avoid detection by T cells, but this leaves the infected cell vulnerable to lysis by natural killer (NK) cells responding to changes in MHC presentation. It is therefore plausible that CrHV2 captured an MHC homolog to disguise infected cells by using a decoy MHC molecule, a strategy that has been demonstrated

for other viruses (42, 43). In this context, the increased purifying selection may act to preserve the recognition of the decoy by NK cells.

Capture-and-replace evolutionary mechanism. We suggest that the selection shift in these host-derived rhadinoviral genes is an adaptive mechanism that falls under the more general evolutionary shortcut strategy, and it will be interesting to know how widely used it is in other herpesviruses, and indeed other viral groups. Many host-derived genes in viruses are lineage specific, and the same genes are frequently targeted for capture independently. For example, the IL-10 gene has been captured repeatedly and independently by a variety of DNA viruses from different donor hosts (3, 4). The relative increase in purifying selection we detected may explain why there is a high turnover of such captured genes. Indeed, all of the immunomodulatory cellular genes we analyzed are limited to one or two viruses, which is an indication of specialized capture in response to a specific host, and this includes rhadinoviral genes that we did not analyze in depth due to low sequence similarity. A possible explanation for this capture-and-replace process is that a specific version of an immunomodulatory host gene is being exploited (which is what drives the purifying selection). However, the efficacy of such a gene will be lost once the fitness landscape shifts or if the virus switches to a new host. In these events, it is simple for the virus to recapture an "updated" host homolog, against which the host target has not actively been evolving to evade.

Purifying selection as a tool among a repertoire of strategies. In contrast to such short-lived gene captures, we also analyzed homologs of genes for FGAM synthase that are present in all of the rhadinoviruses, indicating that they were anciently captured and retained rather than lost and replaced. FGAM is an enzyme involved in purine synthesis and is therefore unlikely to be captured for a transient purpose, like the immune evasion of a particular host with a specific immune strategy. Similarly, homologs of genes for the host complement control protein (CCP) are in all of the rhadinoviruses (except BoHV4) and are capable of evading the complement system (44). Both these captured genes are found in orthologous positions, and our analysis reveals that they have undergone a number of duplications, rearrangements, and losses (Fig. 8).

The fact that these genes are not acquired through the frequent capture-and-replace mode may indicate that they are not evolving in an arms-race framework and therefore remain functionally useful for a longer time. However, in HHV8, a host-derived CCL3-like gene appears to have been duplicated twice since capture, indicating that as for the ancient genes, such as the FGAM synthase gene, this can also occur at shorter evolutionary timescales for lineage-specific genes (Fig. 5C). These homologs appear to have gained some distinct functions but nonetheless functionally overlap in their ability to target the same receptors on Th2 cells, as well as all sharing the capacity to recruit $CD4^+$ and $CD25^+$ T cells to downmodulate the immune response (45). Together these observations indicate that the increased purifying selection we detect in some host-derived rhadinoviral genes is specific to genes that are at the forefront of the genetic arms race and are part of a wider strategy of pathogen mimicry, where genes that are functionally maintained are "perfect mimics" and those that are repurposed are "imperfect mimics" (38). In both cases, however, the mimicry is achieved through post-capture divergent evolution rather than by convergence of nonhomologous genes that evolve host-like functions independently (38).

There is therefore a spectrum of possible postcapture evolutionary trajectories. At one extreme are the fastest-evolving genes that are captured and soon replaced, once their efficacy has expired. On the other end are slower-evolving genes that continue to be effective because they do not engage in arms-race dynamics. We also see intermediate examples in our results, such as CD200, for which the selection shift is not as clear (Fig. 7) and the base composition analysis shows that the viral homolog is between the host and virus in terms of composition (Fig. 9).

Captured cellular genes in viruses are an opportunity to directly examine differences without the confounding variable of different genetic sequences. The transfer of genes

from hosts to viruses offers us a natural experiment in which the genetic sequence is controlled for, allowing us to examine the effects of selection on the same gene (not just similar genes) in vastly different genomes and evolutionary contexts. *dN*/*dS* analyses to compare selection pressures among homologs are most often performed on orthologous sequences, although there are no theoretical obstacles to estimating *dN*/*dS* values for genes that share a common ancestry through HGT (i.e., xenologs). For instance, *dN*/*dS* tests have been used to examine the differences in selection between genes in parasitic plants and their homologs in host plants (46). As with most of the genes in our study, the genes reported in that study were also found to be under stronger purifying selection after HGT. Moreover, such an approach has also been used to investigate the different evolutionary dynamics affecting duplicated genes and horizontally acquired genes in bacteria, revealing that paralogs generally evolve slower than xenologs (47). Selection analyses have also been performed on viral interleukin-10 and host homologs, comparing selection pressures on either side of the capture event (48). It should be noted that such tests would be inappropriate if the similarity between genes was due to convergent evolution, not homology, which is why it is necessary to undertake such comparisons from within a rigorous phylogenetic framework. Indeed, in our analyses, four such genes (Fig. S1) did not yield reliable phylogenetic results and thus were not examined for a selection difference.

While captured genes in viruses have been studied before, our study is a detailed and broad investigation into the mechanisms involved, framed from the perspective of the evolutionary arms race between rhadinoviruses and their hosts. Because the genes being captured were themselves shaped by selective pressure from a variety of viruses, this kind of gene capture is a natural experiment that allows us to study the dynamics of protein evolution on either side of the molecular arms race. In these rhadinoviruses, some of the captured host genes are exploiting the selective consequences of a range of different arms races between the host and many different viruses. Hosts themselves also use this strategy by capturing viral genes that are repurposed to function in antiviral defense. In either case, viruses and hosts are undertaking evolutionary shortcuts to compete in the arms race, in the form of genes that could have emerged only from arms-race dynamics.

MATERIALS AND METHODS

Sequence collation, orthology detection, and alignment. Full-genome alignments for seven rhadinoviruses were downloaded from the NCBI database in GenBank format and aligned using ProgressiveMauve (13) (Fig. 1). A table of all coding sequence accession numbers for each virus was constructed, in which each row represents a group of genes that probably represent orthologs. These orthologies were hypothesized using the results of a BLASTp search of each protein sequence against all other *Rhadinovirus* proteins, as well as an assessment of synteny from whole-genome alignments (see Table S3 in the supplemental material). Each row of likely orthologs was then used as BLASTp and tBLASTn queries (or a group of queries) against the NCBI nr, nt, RefSeq, and wgs sequence databases, with viruses excluded (except for the wgs database, which does not contain viruses) in order to identify potential homologs in nonviral species. Any results with a BLAST score above 40 were manually examined for the potential to be aligned with the viral sequences. We then short-listed genes that exhibited at least 30% identity but excluded those for which homologous sites could be identified for fewer than 50 amino acids. These were normally BLAST alignments that were shorter than 100 amino acids and/or where the identity was scattered along the gene, making it difficult to infer homologous positions between conserved regions that act as alignment anchors. The alignments were constructed using MUSCLE as a starting point, with manual editing guided by the BLAST alignments and visual assessment.

Taxon choice and phylogenetic reconstruction. We performed a phylogenetic analysis for each of the short-listed viral genes to assess their relationship to host homologs. We constructed alignments consisting of a diverse range of mammalian species (GenBank accession numbers are listed in Table S6). Our taxon choice included more primate representatives than representatives of any other order, since 4 of the 7 viruses infect primate hosts and we wished to evaluate the genes' relationships in this part of the tree in finer detail. All phylogenetic reconstruction was performed in parallelized MrBayes (49, 50). We ran two independent Markov chain Monte Carlo (MCMC) chains for 10 million generations to ensure convergence (we used the effective sample size of the posterior probability as a measure of convergence). Due to the high divergence of viral homologs, we favored amino acid alignments to mitigate the effect of highly divergent viral sequences, but nucleotide versions were also used in the instances where poor statistical support was exhibited in the amino acid tree (posterior probabilities below 85% for the nodes relevant to our investigation). The best evolutionary model for each alignment was determined using JModelTest 2 (51) or ProtTest 3 (52), according to the corrected Akaike information criterion. The

tree shown in Fig. 2 was reconstructed from a 266-amino-acid alignment of MHC class 1 genes. The CrHV2 homolog was derived from the three separate predicted ORFs, R1, R2, and R3 (RV138 to -140 in Table S3). For Fig. 3, the IL-17 tree was reconstructed using a 144-amino-acid alignment of 36 mammalian IL-17 genes and a homolog in SaHV2 (RV134 in Table S3). The DHFR tree in Fig. 6A was reconstructed from a 176-amino-acid alignment of DHFR genes from 35 mammalian species and their homologs from SaHV2, MaHV5, and HHV8 (RV10 in Table S3). The CD59 tree in Fig. 5A was reconstructed from a 100-amino-acid alignment of 29 mammalian genes and the SaHV2 homolog (RV136 in Table S3). The C2GnT-M tree in Fig. 5B was reconstructed from a 431-amino-acid alignment of 45 mammalian C2GnT-M genes and the BoHV4 homolog (RV130 in Table S3). Additional taxa that are closely related to *Bos taurus* were added in this case to evaluate the relationship of the viral gene to host homologs in finer detail. The best model for all five of these amino acid-based trees was JTT+G. The CCL3 tree shown in Fig. 5C was constructed using a 93-nucleotide alignment of 35 mammalian CCL3 genes, an MaHV5 homolog, and 3 homologs identified in HHV8 (RV13, RV14, and RV17 in Table S3). The CASP8 tree shown in Fig. S1 was reconstructed from a nucleotide alignment of 36 mammalian genes and their homologs from BoHV4, SaHV2, AtHV3, HHV8, and MaHV5 (RV79 in Table S3). GTR+G was the best evolutionary model for both of these trees. For a number of genes, neither amino acid nor conventional nucleotide models succeeded in reconstructing reliable trees (i.e., there were low posterior probabilities across the tree and/or no convergence of MCMC chains). We therefore implemented the SRD06 model (separate HKY+G models for codon 1 plus codon 2 and codon 3) in the case of the IL-6 tree (Fig. 4A), the CD200 tree (Fig. 7A), and the R1 and acyl-CoA trees (Fig. S1). SRD06 is thought to be more biologically realistic and has been shown to frequently outperform other approaches (53).

Selection analysis. We assessed the effect of viral capture on selection pressure by using a maximum likelihood approach implemented in CODEML. Although such *dN*/*dS* analyses are usually performed using orthologs, there are precedents for the analysis of other types of homologs, such as paralogs and xenologs (46–48, 54, 55). We tested for individual amino acid sites under positive selection in the alignment for eight genes (excluding those for which the precise placement of the viral branch was not statistically robust). The alignment used was the same as that for phylogenetic reconstruction (i.e., with highly divergent/unalignable regions and indels removed). We also used the tree we constructed for the analysis as an input to PAML (software for the analysis of selection under a maximum likelihood framework [56]). At least one positively selected site was identified in all but two of the genes tested (Table S1). For each of the eight genes, we tested whether our results may have been biased by the fact that we chose to include many primate sequences that are very similar, which could have obscured positively evolving sites (Table S5). We therefore repeated the analysis with only five primate sequences and found that this had a small influence on the results, with fewer sites detected in CD59, IL-6, and MHC class II genes (Table S5). We also implemented the branch site test to detect positively evolving sites along the viral branch (or internal viral branch in the trees with multiple viral tips). The goal was to identify the sites that underwent rapid evolutionary change immediately after being captured, but the analysis revealed no such sites. It should be noted that the sensitivity and power of this test are highly dependent on the nature of the data set, requiring an ideal divergence level, alignment length, and sample size (57). In our case, we speculate that the size of the data set may have influenced the negative result, especially since the changes we were trying to detect would have been fleeting in evolutionary terms. It would be interesting for future studies to simulate the ranges of data sizes and types necessary to detect such an event.

We used the branch model test implemented in CODEML, in which an *a priori* model of selection is compared against a null hypothesis. For each gene, we tested the null hypothesis that the *dN*/*dS* ratio is the same across the tree (Table S1) (model = $\omega 0$) and an alternative model in which the viral branch *dN*/*dS* ratio is estimated separately. For three of the genes, there is more than one viral branch for genes that are found in multiple viruses (homologs of CD200, CCL3, and DHFR). We therefore tested three alternative models, allocating separate *dN*/*dS* ratios for the branch internal to the viral group, estimating separate *dN*/*dS* ratios for all viral branches, or designating separate values for the internal branch and viral tips (Table S1). In order to examine the selection dynamics along each gene in finer detail, we developed and implemented a sliding-window modification of the branch test. In our modification, rather than comparing the overall *dN*/*dS* values for the branches under consideration, we analyzed subsections of the alignment, plotting the *dN*/*dS* values along with whether or not a significant difference was exhibited. A custom script was written to run multiple instances of CODEML for a window of 120 bp, with a step size of 3 bp. Both the null and alternative hypotheses were run for each window, and a likelihood ratio test was used to determine significance according to the χ^2 test, with a *P* value threshold of 0.05. This *P* value threshold was used as a proxy for the magnitude of the effect, meant to highlight the windows that exhibited the largest differences in *dN*/*dS* value. It should not be interpreted as an indicator of significance, except in the case of the overall test that identified the general trend. Rather, the sliding-window approach is an exploratory tool that illuminates the heterogeneity of the effect we are detecting. Window sizes of 30, 60, 90, and 150 bp were also attempted, but these gave poorer results—shorter windows produced fewer stretches of statistically significant differences in *dN*/*dS* value, and a larger window resulted in lower resolution for shorter genes. Note that since the same alignment was used for the standard analysis and the sliding-window modification, the *x* axis scales in all charts shown in Fig. 2, 3, 4, 6, and 7 are the same. The scale is also the same along the *y* axis, and the tick spacing is optimized for readability in each case.

Sequence composition analysis, structural modeling, and cophylogeny analysis. For each alignment in our main analysis, we measured the gene lengths, CpG bias, and all 4 mononucleotide and 16 dinucleotide frequencies using a custom script. These variables were used in a principal component

analysis to identify whether a pattern could be detected to distinguish viral genes from their host homologs. We examined the first three principal components for all genes analyzed (Fig. 9), which in each case collectively accounted for >90% of the variance. In addition to the viral gene under investigation and its homologs in host species, we also included the rest of the genes in the viral genome to which the gene belongs to assess their location in the plot relative to the host gene and viral homolog.

We used the Phyre2 Web portal to model the tertiary structure of viral homologs to evaluate the potential functional significance of any sequence divergence since HGT (58). To enable direct comparison with the results of the selection analyses, we used the same amino acid sequence as the PAML input (i.e., with some trimmed regions that could not be aligned to host homologs). We generated an overlap of the model generated with the database template that is most likely a homolog of the query in the UCSF Chimera package (59) (Fig. 2 to 5). In addition, we visually mapped the results of the selection analysis on each residue as a heat map representing *dN*/*dS* estimates. The *dN*/*dS* values are represented along a color spectrum of blue to red, where each increment represents an increase of the *dN*/*dS* ratio of 0.25, with the darkest blue being equivalent to 0 and the darkest red representing the highest estimated value. It should be noted that such modeling approaches infer structure from sequence homology and should be interpreted as a rough guide for how compatible the analyzed regions of a viral homolog are to a reference structure. For this reason, we do not consider the models to be predicted structures of the actual protein. However, by modeling the corresponding homologous regions, this technique allowed us to reveal the structural placement of selected sites rather than only identifying their sequence coordinates. A crucial limitation to this approach is interpreting how/if the unmodeled differences between similar sequences would influence structure, and therefore function. Nonetheless, the approach allows us to draw conclusions from the differences we observe at major structures that are more likely to be modeled correctly.

For the FGAM synthase and CCP homologs, we conducted a cophylogeny analysis using JANE 4.0, which is an event-based method of reconciling tree topologies of hosts and parasites (60). The method works by considering five possible evolutionary events—cospeciation, duplication, loss, host switching, and failure to diverge—and reconstructs possible evolutionary histories that can explain the data with a minimum cost of events. In our analysis, we used the default cost scheme of 0, 1, 2, 1, and 1 for cospeciation, duplication, loss, host switching, and failure to diverge, respectively, and ran an additional analysis that disallowed host switching. Note that in the context of our investigation, the host switching parameter is the equivalent of horizontal gene transfer, since we compared the degrees of congruence of tree topologies between the viral phylogeny and a gene tree rather than a host-parasite tree. We calculated the distribution of costs for random sample solutions for 1,000 samples, using random tip mapping as the randomization method.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/JVI.00359-18.

SUPPLEMENTAL FILE 1, PDF file, 1.2 MB.

REFERENCES

1. Alcami A. 2003. Viral mimicry of cytokines, chemokines and their receptors. Nat Rev Immunol 3:36–50. https://doi.org/10.1038/nri980.
2. Holzerlandt R, Orengo C, Kellam P, Albà MM. 2002. Identification of new herpesvirus gene homologs in the human genome. Genome Res 12:1739–1748. https://doi.org/10.1101/gr.334302.
3. Schönrich G, Abdelaziz MO, Raftery MJ. 2017. Herpesviral capture of immunomodulatory host genes. Virus Genes 53:762–773. https://doi.org/10.1007/s11262-017-1460-0.
4. Ouyang P, Rakus K, van Beurden SJ, Westphal AH, Davison AJ, Gatherer D, Vanderplasschen AF. 2014. IL-10 encoded by viruses: a remarkable example of independent acquisition of a cellular gene by viruses and its subsequent evolution in the viral genome. J Gen Virol 95:245–262. https://doi.org/10.1099/vir.0.058966-0.
5. McGeoch DJ, Davison AJ, Dolan A, Gatherer D, Sevilla-Reyes EE. 2008. Molecular evolution of the Herpesvirales, p 447–475. *In* Domingo E, Parrish CR, Holland JJ (ed), Origin and evolution of viruses. Academic Press, Cambridge, MA.
6. Feschotte C, Gilbert C. 2012. Endogenous viruses: insights into viral evolution and impact on host biology. Nat Rev Genet 13:283–296. https://doi.org/10.1038/nrg3199.
7. Holmes EC. 2011. The evolution of endogenous viral elements. Cell Host Microbe 10:368–377. https://doi.org/10.1016/j.chom.2011.09.002.
8. Katzourakis A, Gifford RJ. 2010. Endogenous viral elements in animal genomes. PLoS Genet 6:e1001191. https://doi.org/10.1371/journal.pgen.1001191.
9. Katzourakis A. 2013. Paleovirology: inferring viral evolution from host genome sequence data. Philos Trans R Soc Lond B Biol Sci 368:20120493. https://doi.org/10.1098/rstb.2012.0493.
10. Aswad A, Katzourakis A. 2012. Paleovirology and virally derived immunity. Trends Ecol Evol 27:627–636. https://doi.org/10.1016/j.tree.2012.07.007.
11. Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351:1083–1087. https://doi.org/10.1126/science.aad5497.
12. Pellett PE, Biggin MD, Barrell B, Roizman B. 1985. Epstein-Barr virus genome may encode a protein showing significant amino acid and predicted secondary structure homology with glycoprotein B of herpes simplex virus 1. J Virol 56:807–813.
13. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403. https://doi.org/10.1101/gr.2289704.
14. Loh J, Zhao G, Nelson CA, Coder P, Droit L, Handley SA, Johnson LS, Vachharajani P, Guzman H, Tesh RB, Wang D, Fremont DH, Virgin HW. 2011. Identification and sequencing of a novel rodent gammaherpesvirus that establishes acute and latent infection in laboratory mice. J Virol 85:2642–2656. https://doi.org/10.1128/JVI.01661-10.
15. Gu C, Wu L, Li X. 2013. IL-17 family: cytokines, receptors and signaling. Cytokine 64:477–485. https://doi.org/10.1016/j.cyto.2013.07.022.
16. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B, Pontius J, Roelke M, Rumpler Y, Schneider MPC, Silva A, O'Brien SJ, Pecon-Slattery J. 2011. A molecular phylogeny of living primates. PLoS Genet 7:e1001342. https://doi.org/10.1371/journal.pgen.1001342.
17. Opazo JC, Wildman DE, Prychitko T, Johnson RM, Goodman M. 2006. Phylogenetic relationships and divergence times among New World monkeys (Platyrrhini, Primates). Mol Phylogenet Evol 40:274–280. https://doi.org/10.1016/j.ympev.2005.11.015.

18. Kishimoto T. 2010. IL-6: from its discovery to clinical applications. Int Immunol 22:347–352. https://doi.org/10.1093/intimm/dxq030.
19. Ehlers B, Dural G, Yasmum N, Lembo T, de Thoisy B, Ryser-Degiorgis M-P, Ulrich RG, McGeoch DJ. 2008. Novel mammalian herpesviruses and lineages within the Gammaherpesvirinae: cospeciation and interspecies transfer. J Virol 82:3509–3516. https://doi.org/10.1128/JVI.02646-07.
20. Markine-Goriaynoff N, Georgin J-P, Goltz M, Zimmermann W, Broll H, Wamwayi HM, Pastoret P-P, Sharp PM, Vanderplasschen A. 2003. The core 2 beta-1,6-*N*-acetylglucosaminyltransferase-mucin encoded by bovine herpesvirus 4 was acquired from an ancestor of the African buffalo. J Virol 77:1784–1792. https://doi.org/10.1128/JVI.77.3.1784-1792.2003.
21. Dewals B, Thirion M, Markine-Goriaynoff N, Gillet L, de Fays K, Minner F, Daix V, Sharp PM, Vanderplasschen A. 2006. Evolution of bovine herpesvirus 4: recombination and transmission between African buffalo and cattle. J Gen Virol 87:1509–1519. https://doi.org/10.1099/vir.0.81757-0.
22. Berg J, Tymoczko J, Stryer L. 2002. Biochemistry, 5th ed. W H Freeman & Co Ltd, New York, NY.
23. Cinquina CC, Grogan E, Sun R, Lin SF, Beardsley GP, Miller G. 2000. Dihydrofolate reductase from Kaposi's sarcoma-associated herpesvirus. Virology 268:201–217. https://doi.org/10.1006/viro.1999.0165.
24. Chung Y-H, Means RE, Choi J-K, Lee B-S, Jung JU. 2002. Kaposi's sarcoma-associated herpesvirus OX2 glycoprotein activates myeloid-lineage cells to induce inflammatory cytokine production. J Virol 76:4688–4698. https://doi.org/10.1128/JVI.76.10.4688-4698.2002.
25. Kirkitadze MD, Barlow PN. 2001. Structure and flexibility of the multiple domain proteins that regulate complement activation. Immunol Rev 180:146–161. https://doi.org/10.1034/j.1600-065X.2001.1800113.x.
26. Rezaee SAR, Cunningham C, Davison AJ, Blackbourn DJ. 2006. Kaposi's sarcoma-associated herpesvirus immune modulation: an overview. J Gen Virol 87:1781–1804. https://doi.org/10.1099/vir.0.81919-0.
27. Lee B-S. 2000. Inhibition of intracellular transport of B cell antigen receptor complexes by Kaposi's sarcoma-associated herpesvirus K1. J Exp Med 192:11–22. https://doi.org/10.1084/jem.192.1.11.
28. Zong JC, Ciufo DM, Alcendor DJ, Wan X, Nicholas J, Browning PJ, Rady PL, Tyring SK, Orenstein JM, Rabkin CS, Su IJ, Powell KF, Croxson M, Foreman KE, Nickoloff BJ, Alkan S, Hayward GS. 1999. High-level variability in the ORF-K1 membrane protein gene at the left end of the Kaposi's sarcoma-associated herpesvirus genome defines four major virus subtypes and multiple variants or clades in different human populations. J Virol 73:4156–4170.
29. Sanjuán R. 2012. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. PLoS Pathog 8:e1002685. https://doi.org/10.1371/journal.ppat.1002685.
30. Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9:267–276. https://doi.org/10.1038/nrg2323.
31. Katzourakis A, Aswad A. 2016. Evolution: endogenous viruses provide shortcuts in antiviral immunity. Curr Biol 26:R427–R429. https://doi.org/10.1016/j.cub.2016.03.072.
32. Holmes EC. 2011. What does virus evolution tell us about virus origins? J Virol 85:5247–5251. https://doi.org/10.1128/JVI.02203-10.
33. Renzette N, Gibson L, Bhattacharjee B, Fisher D, Schleiss MR, Jensen JD, Kowalik TF. 2013. Rapid intrahost evolution of human cytomegalovirus is shaped by demography and positive selection. PLoS Genet 9:e1003735. https://doi.org/10.1371/journal.pgen.1003735.
34. Shackelton LA, Holmes EC. 2004. The evolution of large DNA viruses: combining genomic information of viruses and their hosts. Trends Microbiol 12:458–465. https://doi.org/10.1016/j.tim.2004.08.005.
35. Elde NC, Child SJ, Eickbush MT, Kitzman JO, Rogers KS, Shendure J, Geballe AP, Malik HS. 2012. Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. Cell 150:831–841. https://doi.org/10.1016/j.cell.2012.05.049.
36. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. Nat Rev Genet 8:857–868.
37. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. Science 312:1614–1620. https://doi.org/10.1126/science.1124309.
38. Elde NC, Malik HS. 2009. The evolutionary conundrum of pathogen mimicry. Nat Rev Microbiol 7:787–797. https://doi.org/10.1038/nrmicro2222.
39. Schmid K, Yang Z. 2008. The trouble with sliding windows and the selective pressure in BRCA1. PLoS One 3:e3746. https://doi.org/10.1371/journal.pone.0003746.
40. Bernatchez L, Landry C. 2003. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J Evol Biol 16:363–377. https://doi.org/10.1046/j.1420-9101.2003.00531.x.
41. Spurgin LG, Richardson DS. 2010. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc Biol Sci 277:979–988. https://doi.org/10.1098/rspb.2009.2084.
42. Babić M, Krmpotić A, Jonjić S. 2011. All is fair in virus-host interactions: NK cells and cytomegalovirus. Trends Mol Med 17:677–685. https://doi.org/10.1016/j.molmed.2011.07.003.
43. Pyzik M, Dumaine A, Dumaine AA, Charbonneau B, Fodil-Cornu N, Jonjic S, Vidal SM. 2014. Viral MHC class I-like molecule allows evasion of NK cell effector responses in vivo. J Immunol 193:6061–6069. https://doi.org/10.4049/jimmunol.1401386.
44. Lambris JD, Ricklin D, Geisbrecht BV. 2008. Complement evasion by human pathogens. Nat Rev Microbiol 6:132–142. https://doi.org/10.1038/nrmicro1824.
45. Nicholas J. 2005. Human gammaherpesvirus cytokines and chemokine receptors. J Interferon Cytokine Res 25:373–383. https://doi.org/10.1089/jir.2005.25.373.
46. Yang Z, Zhang Y, Wafula EK, Honaas LA, Ralph PE, Jones S, Clarke CR, Liu S, Su C, Zhang H, Altman NS, Schuster SC, Timko MP, Yoder JI, Westwood JH, dePamphilis CW. 2016. Horizontal gene transfer is more frequent with increased heterotrophy and contributes to parasite adaptation. Proc Natl Acad Sci U S A 113:E7010–E7019. https://doi.org/10.1073/pnas.1608765113.
47. Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet 7:e1001284. https://doi.org/10.1371/journal.pgen.1001284.
48. Jayawardane G, Russell GC, Thomson J, Deane D, Cox H, Gatherer D, Ackermann M, Haig DM, Stewart JP. 2008. A captured viral interleukin 10 gene with cellular exon structure. J Gen Virol 89:2447–2455. https://doi.org/10.1099/vir.0.2008/001743-0.
49. Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574. https://doi.org/10.1093/bioinformatics/btg180.
50. Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20:407–415. https://doi.org/10.1093/bioinformatics/btg427.
51. Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772. https://doi.org/10.1038/nmeth.2109.
52. Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165. https://doi.org/10.1093/bioinformatics/btr088.
53. Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol 23:7–9. https://doi.org/10.1093/molbev/msj021.
54. Bustos O, Naik S, Ayers G, Casola C, Perez-Lamigueiro MA, Chippindale PT, Pritham EJ, de la Casa-Esperón E. 2009. Evolution of the Schlafen genes, a gene family associated with embryonic lethality, meiotic drive, immune processes and orthopoxvirus virulence. Gene 447:1–11. https://doi.org/10.1016/j.gene.2009.07.006.
55. Pegueroles C, Laurie S, Albà MM. 2013. Accelerated evolution after gene duplication: a time-dependent process affecting just one copy. Mol Biol Evol 30:1830–1842. https://doi.org/10.1093/molbev/mst083.
56. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591. https://doi.org/10.1093/molbev/msm088.
57. Yang Z, dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. Mol Biol Evol 28:1217–1228. https://doi.org/10.1093/molbev/msq303.
58. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10:845–858. https://doi.org/10.1038/nprot.2015.053.
59. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612. https://doi.org/10.1002/jcc.20084.
60. Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R. 2010. Jane: a new tool for the cophylogeny reconstruction problem. Algorithms Mol Biol 5:16. https://doi.org/10.1186/1748-7188-5-16.