

1 Cell-derived viral genes evolve under stronger purifying selection in  
2 Rhadinoviruses

3 Running title: increased purifying selection in host derived viral  
4 genes

5

6

7

8 Amr Aswad<sup>1</sup> and Aris Katzourakis<sup>1\*</sup>

9

10 <sup>1</sup>Zoology Department,

11 University of Oxford,

12 South Parks Road

13 Oxford, Oxfordshire

14 Ox1 3PS

15

16 \*Corresponding author

17 aris.katzourakis@zoo.ox.ac.uk

18

## 19 Abstract

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

## 34 Importance

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

## 42 Introduction

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

53 Like other herpesviruses (and indeed many large dsDNA viruses), rhadinoviruses can adaptively  
54 capture host genes that increase their fitness through horizontal gene transfer (HGT). Amidst 'core  
55 blocks' of genes that are conserved across herpesviruses, lineage-specific genes include captured  
56 cellular immunomodulatory genes such as cytokines, chemokines and interferon regulatory factors  
57 (1–3). These are the same genes that evolve at a particularly high rate in vertebrates, driven by the  
58 selection pressure imposed by viruses.

59 The viral evolutionary strategy of cellular gene capture is so successful that some genes such as  
60 interleukin 10 have been repeatedly captured multiple times independently in different viral groups  
61 (3, 4). Rhadinoviruses are an ideal study group to systematically reconstruct the evolutionary  
62 history of cell-derived genes, because they are known to frequently capture host genes and there  
63 are seven well-annotated genomes with lineage-specific cell-derived ORFs (5). Moreover, the  
64 genus includes *Human herpesvirus 8* (HHV8) that is the etiological agent of Kaposi's sarcoma, for  
65 which ORFs with sequence similarity to human genes have already been identified (2).

66 Gene-flow between viruses and their hosts occurs in both directions, with numerous examples of  
67 viral genomes becoming heritably integrated in the genomes of their hosts. Known as endogenous  
68 viral elements (EVEs), (6–9). they are usually non-functional relics, but can play a beneficial role in

69 the host genome, including as antiviral defense genes, known as EVE-derived immunity genes  
70 (EDIs) (6, 10, 11). The capture of these EDIs is a mechanism employed by hosts as part of the  
71 evolutionary arms race to defend against viruses, just as viruses capture cellular genes to evade  
72 host immunity.

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

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

## 95 Results

96 We constructed a whole genome alignment of seven rhadinoviruses using Mauve (14) to structure  
97 the investigation according to the locus of each putative gene capture (Fig 1). We focused on  
98 genes outside the conserved core blocks, which we know to be shared across all herpesviruses,  
99 and arbitrarily designated each of the remaining variable regions A-C (Fig 1). Each gene was  
100 matched to its likely homolog in other rhadinoviruses by searching for sequence similarity using  
101 BLASTp with a minimum alignment score threshold of 50. Sequence-similar rhadinoviral genes  
102 were considered syntenic if they shared similar flanking genes in the equivalent region of the  
103 genome. BLAST was also used for each gene to identify similarity to host sequences, which is an  
104 indication that the viral sequence could have been derived by HGT. We excluded genes that were  
105 too divergent to align to similar host genes with confidence, i.e. where homologous sites could be  
106 hypothesized for fewer than 50 amino acids. This resulted in a shortlist of 13 genes that we  
107 considered for further phylogenetic and evolutionary analysis.

### 108 Major histocompatibility complex, IL-6 and IL-17

109 The first three genes in *Cricetid gammaherpesvirus 2* (CrHV2), designated as R1, R2 and R3,  
110 show sequence similarity to major histocompatibility complex (MHC) genes (Fig 1), and were  
111 aligned in the initial characterization of the viral genome with human HLA-A2 and mouse H-2K<sup>b</sup>  
112 (15). In our analysis of these genes, rather than treating them as separate genes, we have  
113 determined that they are in fact similar to different regions of the MHC gene in a co-linear manner.  
114 This lead us to the conclusion that they are derived from the capture of a single gene, and it was  
115 possible to align a concatenation of all three ORFs to a range of MHC homologs. We  
116 reconstructed a phylogeny from a diverse set of mammalian hosts, revealing that CrHV2 R1-3 are  
117 most closely related to the mouse and rat homologs (Fig 2A). Along with the fact that CrHV2 was  
118 isolated from the pygmy rice rat (*Oligoryzomys microtis*), the phylogenetic similarity to mouse and  
119 rat (*Mus musculus* and *Rattus norvegicus*, respectively) is consistent with an acquisition from a  
120 rodent host (Fig 2A). However, none of the mammalian MHCs we identified possess an exon  
121 structure that corresponded to the three CrHV2 ORFs, suggesting the rodent in question has not  
122 been sequenced, and may or may not be extant.

123 Interleukins (ILs) are a large group of cytokines, of which genes from several different families  
124 have been captured by rhadinoviruses. IL-17 is a family of 6 types of interleukin involved in the  
125 inflammatory response (IL-17A to F) (16). *Saimirine gammaherpesvirus 2* (SaHV2) acquired a  
126 copy of IL-17A, and the fact that this gene is not present in *Ateline gammaherpesvirus 3* (AtHV3)  
127 suggests that the capture event occurred after their speciation. Phylogenetic analysis is consistent  
128 with this hypothesis (Fig 3A), since despite the long branch of SaHV2 IL-17A, the gene groups  
129 robustly with the IL-17 genes of squirrel monkey (*S. boliviensis*) and marmoset (*C. jacchus*), which  
130 speciated after the divergence of their common ancestor from spider monkeys (*Ateles sp.*). This  
131 could mean that the gene was acquired by the SaHV2 ancestor after the squirrel monkey-spider  
132 monkey split, but before the squirrel monkey-marmoset split at ~20 million years (17). This result  
133 suggests that the capture event may have occurred in a small 3 million year window, when new  
134 world monkeys were undergoing a major radiation that separated the major clades (*Atelidae*,  
135 *Cebidae* & *Pitheciidae*) (18).

136 The genomes of HHV8 and *Macacine gammaherpesvirus 5* (MaHV5) contain homologs of IL-6,  
137 which is a secreted cytokine known to have a variety of functions, ranging from roles in  
138 immunomodulation to oncogenesis (19). The phylogenetic analysis of these genes revealed the  
139 sequences are embedded within a clade that includes mouse, rat and tree shrew (Fig 4A). This is  
140 surprising given that both HHV8 and MaHV5 are primate-infecting viruses, which suggests that the  
141 gene may have been acquired from a non-primate host, and this is consistent with a history of  
142 cross-species transmission in rhadinoviruses (13).

#### 143 **CD59 glycoprotein & Core 2 $\beta$ -1,6-N-acetylglucosaminyltransferase-mucin (C2GnT-** 144 **M)**

145 SaHV2 encodes a homolog of the mammalian cell-surface glycoprotein CD59, first identified  
146 during the genome-sequencing project of the virus in 1992. The results of our phylogenetic  
147 analysis demonstrate that the gene was acquired through HGT, since the viral homolog resolves  
148 next to *Saimiri boliviensis* with a high posterior probability to the exclusion of *Callithrix jacchus*  
149 CD59 (Fig 5A). We can therefore conclude with confidence that the gene was captured after the

150 *Callithrix-Saimiri* divergence (~20mya (17)) from either the current host (*Saimiri*) or a closely  
151 related monkey that may be susceptible to infection. This is also supported by the fact that the  
152 gene is absent in AtHV3, the most closely related virus to SaHV2.

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

### 160 **Dihydrofolate reductase, CD200 immunoglobulin & chemokine C-C motif ligand 3** 161 **(CCL3)**

162 Several of the host-derived genes investigated were found in multiple viruses, suggesting that they  
163 were acquired before the viruses diverged. The genomes of MaHV5 and HHV8 also include ORFs  
164 with similarity to the small cellular C-C motif chemokine ligand CCL3 that plays a role in the  
165 inflammatory response. Three such genes were identified in HHV8, and phylogenetic  
166 reconstruction places the clade of viral genes within primates with over 80 probability, but their  
167 exact position is not well supported (Fig 5C). The fact that the HHV8 genes group together to the  
168 exclusion of the MaHV5 copy with 97 probability indicates that they were duplicated within the  
169 herpesvirus after capture.

170 Three rhadinoviruses contain a homolog of dihydrofolate reductase (DHFR), which is a universal  
171 cellular enzyme that catalyses the conversion of dihydrofolate to tetrahydrofolate (22).  
172 Tetrahydrofolate is required for purine and thymidylc acid synthesis, as well as synthesis of amino  
173 acids such as glycine and cysteine (22). Although the gene is situated in variable region A for  
174 MaHV5 and SaHV2, the HHV8 DHFR homolog is located downstream of this in variable region B.  
175 If all three genes originate from the same capture event, then the DHFR in HHV8 must have either  
176 been moved through recombination between variable region A and B, or else was duplicated and

177 then lost in region A. However, there is a 23 amino acid long carboxyl-terminus region in HHV8  
178 DHFR that is absent in MaHV5 and SaHV2, a difference that has been pointed out as supportive of  
179 a separate capture scenario to explain the distinct locus (23). Although the tree in this study shows  
180 that the genes cluster together with 98 posterior probability, it is not possible to rule out a separate  
181 capture since intermediate lineages may not have been sampled (Fig 6A).

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

#### 189 **Phosphoribosylformylglycinamide synthase & complement control protein**

190 Although this study is not focused on genes that were likely captured before the emergence of  
191 rhadinoviruses, we examined the evolution of two such genes due to their extensive post capture  
192 duplication and rearrangement in rhadinoviruses. Phosphoribosylformylglycinamide synthase  
193 (FGAMS) is a purine synthesis enzyme, which was captured early in the evolution of  
194 gammaherpesviruses, being shared by all the viruses in the group. Among rhadinoviruses  
195 however, a number of duplications, deletions and translocations have occurred, though have not  
196 been fully investigated. We conducted a phylogenetic reconstruction of available  
197 gammaherpesvirus FGAMS genes to trace the series of genomic changes that lead to the current  
198 state of *Rhadinovirus* FGAMS (Fig 8A & 8B). Consistent with previous observations, the tree  
199 shows that after the ancestral capture, two initial duplication event occurred resulting in three  
200 copies in the ancestor of macaviruses, percaviruses and rhadinoviruses. Both new copies appear  
201 to have been lost in the MaHV5-HHV8 ancestor, but duplicated a second time in the ancestor of  
202 CrHV2, *Murid gammaherpesvirus 4* (MuHV4), *Wood mouse herpesvirus* (WMHV). These rodent  
203 viruses also appear to have lost the ancestral FGAMS, since their copies are more closely related  
204 to one of the copies that arose in the ancient double duplication. This suggests that a translocation



205 event may have occurred at the time of this secondary event. This second duplicate was itself  
206 duplicated in the MuHV4-WMHV ancestor, since both genomes contain a third copy that is most  
207 closely related to it.

208 A homolog of the host gene that encodes the complement control protein (CCP) is found in all  
209 rhadinoviruses except BoHV4. As with several genes of the complement system (a subset of  
210 innate immunity), CCP contains modular sushi repeats, which are conserved ~60 amino acid  
211 domains that exist in a range of complement and adhesion proteins (25). In rhadinoviruses, the  
212 number of sushi domains is variable in the different homologs (between 4 and 8). The rhadinoviral  
213 homologs are most similar at the sequence level to the conserved mammalian complement  
214 component 4 binding protein (determined by BLAST), but could not be aligned with confidence due  
215 to low sequence identity. Furthermore, it is reasonable to assume that the CCP genes are  
216 orthologs because they are all found at roughly the same locus and are partially syntenic (Fig 1).  
217 HHV8, MuHV4 and SaHV2 each have one gene with similarity to CCP containing 4 sushi domains,  
218 but none of the most closely related viruses to each of these viruses possess a CCP ortholog with  
219 the same number of sushi domains. Based on a high posterior probability grouping the genes in  
220 the phylogeny, we can conclude that they originate from a single capture of a gene that likely had 4  
221 sushi domains, based on the domain organization of a similar CCP host gene (CD46). Based on  
222 these conclusions, the AtHV3 appears to have lost a single sushi domain since it diverged from  
223 SaHV2. In MaHV5, an apparent duplication has resulted in a gene that is twice as long as the  
224 HHV8 copy (with 8 sushi domains). A similar duplication seems to have occurred in CrHV2  
225 compared to MuHV4, but in this case they are two separate ORFs separated by several genes and  
226 each missing two sushi domains.

227 To investigate the possible evolutionary history of duplication and loss of CCP genes, we  
228 conducted a co-phylogeny analysis using JANE 4.0 (Fig 8D & 8E). This allowed us to reconstruct  
229 the possible series of events based on the respective phylogenetic relationships among the genes  
230 and the relative topology of the gene tree to the virus tree. This revealed one most parsimonious  
231 solution when HGT between the viruses is permitted, and one solution where no HGT is allowed.  
232 In both instances, we used the default cost settings for duplication, loss and failure to diverge.

233 Under the latter scenario, only SaHV2 and AtHV3 have retained the ancestrally captured gene,  
234 which was replaced with subsequent duplicates in all other viruses. The first duplication gave rise  
235 to the 5' end of the MaHV4 CCP as well as the copy in CrHV2 and MuHV4. This copy underwent a  
236 further two rounds of duplication leading to the HHV8 copy and the 3' end of the MaHV5 gene. In  
237 CrHV2, a lineage-specific duplication gave rise to the second ORF, after which both genes lost 2  
238 sushi domains. Alternatively, under a model where HGT is permitted, the analysis reveals that the  
239 ancestral copy gave rise to the contemporary gene in SaHV2, AtHV3, HHV8 and the 5' end of the  
240 MaHV5 gene. The 3' end in MaHV5 CCP was the result of a duplication and transfer of a copy  
241 from HHV8, which itself also transferred to the MuHV4-CrHV2 ancestor. Thereafter, the extra copy  
242 in CrHV2 originates from a similar duplication and transfer from MuHV4.

#### 243 **Other genes**

244 Some non-core genes exhibited very low similarity to cellular genes, and as such could not be  
245 used as part of a phylogenetic analysis to determine the host group of origin with statistical  
246 confidence. For example, all of the rhadinoviruses apart from MuHV4 and CrHV2 encode a  
247 CASP8-like gene, which in hosts is a protease involved in programmed cell death. Although the  
248 phylogeny was inconclusive with regard to the particular host lineage of origin, it revealed that the  
249 gene does group within mammalian homologs with 93 probability (Fig S1A). Similarly, HHV8 and  
250 MaHV5 share a number of interferon-like genes that were probably acquired after their divergence  
251 from other rhadinoviruses. Interestingly, these genes are located within a single stretch amidst  
252 core block 4 that is otherwise syntenic in nearly all herpesviruses (Fig 1). There are also genes  
253 that could be aligned to host homologs, but phylogenetic analysis was inconclusive. For example,  
254 the BoHV4 genome includes a small hypothetical open reading frame with similarity to Acyl-CoA  
255 thioesterase, which is a family of cellular genes that encode enzymes involved in metabolism (Fig  
256 S1B).

257 HHV8, MaHV5, SaHV2 and AtHV3 all possess another nucleotide metabolism gene - thymidilate  
258 synthase (TS). It is in the same location for MAHV5 and HHV8 but situated near the right terminal  
259 of the genome for SaHV2 & AtHV3 (Fig 1). Although the phylogeny exhibited poor support for most  
260 nodes, the tree reveals that the pair of TS genes in SaHV2 and AtHV3 have much shorter

261 branches than the HHV8/MaHV5 genes, suggesting that they originate from a distinct and more  
262 recent capture (Fig S1C). The fact that the pairs of TS genes are at opposite locations of their  
263 respective genomes also supports this scenario.

264 The leftmost HHV8 gene K1, encodes a transformation-associated glycoprotein that performs roles  
265 in signal transduction, possibly involved in immunomodulation and perpetuating cell survival (26).  
266 Based on a BLAST search against herpesviruses, K1 does not appear to have homologs in other  
267 genomes, except for very weak sequence similarity to MaHV5 R1 (28% amino acid identity, ~50%  
268 query coverage). While we cannot identify host sequences with good similarity to K1, the R1 gene  
269 of MaHV5 is most similar to the N-terminus of the low affinity receptor for the FC fragment of  
270 immunoglobulin G in primates (31% amino acid identity to *Callithrix jacchus*, 41% coverage). It is  
271 possible that this gene is the original host homolog of the viral ORF, which is consistent with  
272 functional studies of K1, which is similar to R1 (albeit distantly). It has been shown that K1 can  
273 cause the down regulation of the B cell antigen receptor complex (BCR), which is known to occur  
274 via the interaction of the N-terminus with the u chain of BCR (27). This would explain why  
275 sequence conservation has been maintained at the N-terminus.

276 We were unable to robustly estimate a phylogeny of the R1 gene with host homologs, probably  
277 due to the effects of long-branch attraction (where a high number of mutations can increase the  
278 chance of spurious branch placement). Nonetheless, the MaHV5 gene groups with placental  
279 mammals with high posterior probability (Fig S1D). K1 and R1 are situated in the same genomic  
280 position and both contain immunoglobulin-like domains. This synteny could be interpreted as  
281 evidence that the genes are highly divergent orthologs that were acquired from the ancestral  
282 primate host, but we cannot rule out the scenario that they are independent acquisitions. The latter  
283 possibility is supported by their extreme divergence from one another compared to the rest of the  
284 genomes (which are otherwise easily aligned). Even among HHV8 K1 genes, the sequence  
285 diversity is extremely high among different strains suggesting that there is a mechanism of  
286 diversifying selection (28), and the same evolutionary pressure can be invoked to explain the  
287 repeated capture of the same gene. An interesting possibility in the case of K1/R1 is that the

288 variability could have been selected in order to mimic or avoid a particular HLA/MHC repertoire  
289 (28).

## 290 **Post-capture evolutionary changes: analysis of selection, sequence composition &** 291 **structural modeling**

292 We determined that the selection pressure imposed on viral genes was significantly different to  
293 that of their cellular homolog (likelihood ratio test,  $p \leq 0.05$ ) in seven of the eight genes tested  
294 (C2GnT-M was the exception). Using the branch-model implemented in codeml, we found that the  
295 viral genes exhibit lower values of dN/dS compared to their cellular counterparts (table S1). The  
296 ratio of nonsynonymous (dN) to synonymous (dS) reflects the balance between neutral,  
297 deleterious or beneficial mutations. Values significantly lower than 1 indicate the presence of  
298 purifying selection, whereas a dN/dS above 1 indicate that beneficial changes are likely being  
299 maintained (i.e. positive selection). For genes where multiple viruses contained a homolog (DHFR,  
300 CD200 and CCL3), we find that the drop in dN/dS is evident whether the model tested  
301 implemented separate estimates for the viral stem and crown group branches or whether both  
302 were designated a single estimate.

303 We also identified between 1-22 sites under positive selection in 6/8 of the genes tested (table S1).  
304 Since the test of positive selection considers the whole alignment for all taxa, the results primarily  
305 reflect the evolutionary history of the host genes, since they represent the majority of sequences in  
306 the data. This indicates that the genes are evolving adaptively in host genomes, and are likely to  
307 be an underestimate of the true number of positively selected sites due to the limitations of  
308 selection analyses (see methods). We therefore tested whether the overall dN/dS reduction in viral  
309 branches according to the branch model could be due to an artifact of these positively evolving  
310 sites in the host that were biasing the dN/dS estimates. We implemented a modification to the  
311 branch-model where these sites are excluded, and find that the drop in dN/dS is still significant  
312 albeit with a slightly lower magnitude (table S2).

313 As well as considering the possible biasing effect of sites evolving under positive selection, we  
314 wanted to explore how different regions of a gene were contributing to the findings of overall dN/dS

315 reduction. We therefore implemented a sliding-window modification of the branch model, which  
316 allows us to consider whether subsections of the gene would return a significant reduction in  
317 dN/dS. In this analysis, the p-value is being used as a proxy for the magnitude of the effect and  
318 does not therefore require correction for multiple testing. Interestingly, we find that the difference in  
319 dN/dS between cellular and viral counterparts is non-uniform along the length of the genes. In the  
320 case of MHC, the test reveals that the viral homolog is only significantly lower in the regions that  
321 correspond to the alpha-1 and alpha-2 domains (Fig 2B). Together these form the antigen binding  
322 cleft, which is where the majority of positively selected sites were detected. Moreover, the viral  
323 dN/dS values within the alpha-1/2 region are lower than the values of the alpha-3 region where the  
324 virus and cell dN/dS cannot be statistically distinguished (Fig 2B).

325 In IL-17A, we find that the significant difference in dN/dS is localized to the conserved IL-17 domain  
326 and signal peptide, and for both IL-6 and DHFR, the drop in dN/dS was limited to roughly half of  
327 the gene (Fig 4B & 6B, respectively). Since a homolog of DHFR is found in more than one  
328 rhadinovirus, the sliding window branch test also revealed a small stretch of significantly different  
329 dN/dS for host branches, viral tips and the viral stem branch (Fig 6B). For the CD200 homolog, we  
330 find that several regions across the gene that exhibit a significant difference in dN/dS between host  
331 and viral branches. Similarly to DHFR, the 3-model branch-test also revealed multiple regions with  
332 a different dN/dS in the viral stem branch (Fig 7B). Unlike the rest of the genes analysed however,  
333 the CD200 analysis shows that the differences in dN/dS are not contiguous, but rather  
334 interspersed between regions where a significant difference could not be detected between virus  
335 and host values. Furthermore, the sliding window test also revealed that the genes are evolving  
336 under comparatively higher dN/dS in the virus, with peaks at 3 different regions (Fig 7B).

### 337 **Sequence composition & structural modeling**

338 We further investigated the differences between viral genes and their host homologs by comparing  
339 nucleotide frequencies (mono- & dinucleotide), CpG bias and gene length in a principal  
340 components analysis. As was the case for the branch-site analysis, the results showed that there  
341 was a substantial difference for all the genes except for the C2GnT-M homologue (Fig 9). We also  
342 included all the genes in each viral genome in the analysis. This revealed that the sequence

343 composition of the captured viral genes was moving away from that of the host from which they  
344 were derived towards the characteristics of the virus genome (Fig 9). In contrast to this, structural  
345 modeling of the most conserved regions of viral homologs shows that the sequences are  
346 compatible with the structure of their host homolog (Figs 2-6). Furthermore, by mapping the  
347 selected sites onto the modeled structure, we are able to corroborate their functional relevance,  
348 and minimize the possibility that the results of the selection analysis are random false positives.  
349 This is particularly striking in the case of the MHC homolog in CrHV2, where the majority of  
350 selected sites are localized to the antigen binding cleft (Fig 2D).

## 351 Discussion

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

358 This high evolutionary rate is reflected in the long branches of rhadinoviral homologs in  
359 phylogenies. Moreover, selective forces have driven the base composition away from that of the  
360 hosts (fig. 9). We are nonetheless able to identify the host lineage of origin for 8 rhadinoviral  
361 genes. For example, the trees in figure 6 and 7 both indicate that the cellular homologs originate  
362 from primates, and more specifically, the CD59 homolog in SaHV2 derived from *Saimiri* sp. with a  
363 posterior probability of 97 (Fig 5A).

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

370 Conservation of function via higher selective constraints is supported by our structural modeling  
371 analyses that revealed that conserved domains are maintained, (i.e. structural similarity is not  
372 simply due to recent acquisition.) While such modeling approaches are not always accurate  
373 predictions of the true protein structure, it demonstrates that the selective forces have not  
374 significantly influenced secondary or tertiary structure. Furthermore, by mapping the changes in  
375 dN/dS to the structural models, we can reveal how the patterns of selection pressure correspond to  
376 the different regions of the protein in 3D space (rather than only considering their location along  
377 the sequence).

### 378 **Gene capture as an evolutionary shortcut**

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

### 392 **Purifying selection in captured genes as one weapon in the evolutionary arms race**

393 The selection tests we developed demonstrate that the rhadinoviral strategy of preservation of  
394 function is achieved through a relative drop in selection pressure (Figs 2-4, 6), in spite of the fact  
395 that host immune genes evolve at high rates compared to other host genes (36, 37). A possible  
396 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 that they need to evade - presumably the host species that it came from, 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 using the standard branch test, we use a sliding window that implements the likelihood ratio test only as a *post hoc* exploratory tool. This allows 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, in IL-6, the largest difference in dN/dS is localized to the first half of the sequence (Fig 4). This could help 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 could be useful to a virus (19).

The adaptive nature of this selection shift is clearest in the IL-17 and MHC homologs (Fig 1 & 3, respectively), where the drop of dN/dS 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 over-represented in the alignment (Fig 1, table S1). Vertebrate MHCs are among the most rapidly evolving vertebrate genes, and there are multiple hypotheses that together explain their paradoxical levels of genetic diversity (40, 41). MHCs function by presenting bound antigens to T-cells to elicit an immune response, and the high diversity of MHCs is driven by the evolutionary pressure to recognize a wide range of antigens (41). A counterstrategy employed by viruses is the down regulation 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 homologue to disguise infected cells using a



424 'decoy' MHC, a strategy that has been demonstrated in other viruses (42, 43). In this context, the  
425 increased purifying selection may act to preserve the recognition of the decoy by NK cells.

#### 426 **A capture-and-replace evolutionary mechanism**

427 We suggest that the selection shift in these host-derived rhadinoviral genes is an adaptive  
428 mechanism that falls under the more general evolutionary short-cut strategy, and it will be  
429 interesting to know how widely used it is in other herpesviruses and indeed other viral groups.  
430 Many host-derived genes in viruses are lineage specific, and the same genes are frequently  
431 targeted for capture independently. For example, IL-10 has been repeatedly captured  
432 independently by a variety of DNA viruses from different donor hosts (3, 4). The relative drop in  
433 selection pressure we detected could explain why there is a high turnover of such captured genes.  
434 Indeed all of the immunomodulatory cellular genes we analyzed are limited to one or two viruses,  
435 which is an indication of specialized capture in response to a specific host, and this includes  
436 rhadinoviral genes that we did not analyze in-depth due to low sequence similarity (34, 44). A  
437 possible explanation for this capture-and-replace process is that a specific 'version' of an  
438 immunomodulatory host gene is being exploited (which is what drives the purifying selection).  
439 However the efficacy of such a gene will be lost once the fitness landscape shifts away, or if the  
440 virus switches to a new host. In these events, it is simple for the virus to re-capture an 'updated'  
441 host homolog, to which the host target has not been actively evolving to evade.

#### 442 **Purifying selection as a tool among a repertoire of strategies.**

443 In contrast to such short-lived gene captures, we also analysed homologs of genes for FGAM-  
444 synthase that are present in all of the rhadinoviruses, indicating that they were anciently captured  
445 and retained rather than lost and replaced. FGAM is an enzyme involved in purine synthesis and is  
446 therefore unlikely to be captured for a transient purpose such as immune evasion of a particular  
447 host with a specific immune strategy. Similarly, homologs of host complement control protein  
448 (CCP) are in all of the rhadinoviruses (except BoHV4) and are capable of evading the complement  
449 system (45). Both these captured genes are found in orthologous positions, and our analysis  
450 reveals that they have undergone a number of duplications, rearrangements and losses (Fig 8).

451 The fact that these genes are not acquired through the frequent capture-and-replace mode could  
452 indicate that they are not evolving in an arms-race framework and therefore remain functionally  
453 useful for a longer time. However, In HHV8, a host-derived CCL3-like gene appears to have  
454 duplicated twice since capture, indicating that like those ancient genes such as FGAM-synthase,  
455 this can also occur at shorter evolutionary timescales for lineage specific genes (Fig 5C). These  
456 homologues appear to have gained some distinct functions, but nonetheless functionally overlap in  
457 their ability to target the same receptors on Th2 cells, as well as all sharing the capacity to recruit  
458 CD4<sup>+</sup> and CD25<sup>+</sup> T-cells to down-modulate the immune response (46). Together, this indicates that  
459 the increased purifying selection we detect in some host-derived rhadinoviral genes is specific to  
460 genes that are at the forefront of the genetic arms-race, and are part of a wider strategy of  
461 pathogen mimicry, where genes that are functionally maintained are so-called 'perfect mimics'  
462 while those that are re-purposed are 'imperfect mimics' (38). In both cases however, the mimicry is  
463 achieved through post-capture divergent evolution, rather than convergence of non-homologous  
464 genes that evolve host-like functions independently (38).

465 There is therefore a spectrum of possible post-capture evolutionary trajectories. At one extreme  
466 are the fastest evolving genes that are captured and soon replaced once their efficacy has reached  
467 its limit. On the other end, are slower genes that continue to be effective because they do not  
468 engage in arms-race dynamics. We also see intermediate examples in our result, such as CD200  
469 where the selection shift is not as clear (Fig 7) and the base composition analysis shows that the  
470 viral homolog is between host and virus composition (Fig 9).

471 Captured cell genes in viruses are an opportunity to directly examine differences without the  
472 confounding variable of different genetic sequences. The transfer of genes from host to viruses  
473 offers us a natural experiment where the genetic sequence is controlled for, allowing us to examine  
474 the effects of selection on the same gene (not just similar) in vastly different genomes and  
475 evolutionary contexts. Using dN/dS analyses to compare selection pressure among homologs is  
476 most often performed on orthologous sequences, although there are no theoretical obstacles to  
477 estimating dN/dS on genes that share common ancestry through HGT (i.e. xenologs). For  
478 instance, dN/dS tests have been used to examine the differences in selection between genes in

479 parasitic plants and their homologs in host plants (47). As with most of the genes in our study, the  
480 genes in this paper were also found to be under stronger purifying selection after HGT. Moreover,  
481 such an approach has also been used to investigate the different evolutionary dynamics affecting  
482 duplicated genes and horizontally acquired genes in bacteria, revealing that paralogs generally  
483 evolve slower than xenologs (48). Selection analyses have also been performed on viral  
484 interleukin-10 and host homologs, comparing selection on either side of the capture event (49). It  
485 should be noted that such tests would be inappropriate, if the similarity between genes was due to  
486 convergent evolution, not homology, which is why it is necessary to undertake a such comparisons  
487 from within a rigorous phylogenetic framework. Indeed in our analyses, four such genes (Fig S1)  
488 did not yield reliable phylogenetic results and so were not examined for the selection difference.

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

## 500 **Materials and Methods**

### 501 **Sequence collation, orthology detection & alignment**

502 Full genome alignments for seven rhadinoviruses were downloaded from NCBI in genbank format  
503 and aligned using ProgressiveMauve (14) (Fig 1). A table of all coding sequence accession IDs  
504 from each virus was constructed where each row represents a group of genes that probably  
505 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 (table S3). Each row of likely orthologs were then used as a BLASTp and tBLASTn query (or group of queries) against the NCBI sequence databases nr, nt, RefSeq and wgs with viruses excluded (except wgs, which does not contain viruses) in order to identify potential homologs in non-viral species. Any results above a BLAST score of 40 were manually examined for their potential to be aligned with the viral sequences. We then shortlisted genes that exhibited at least 30% identity, but excluded those for which homologous sites could only be identified for fewer than 50. 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 & phylogenetic reconstruction**

We performed a phylogenetic analysis for each of the shortlisted viral genes to assess their relationship to host homologs. We constructed alignments consisting of a diverse range of mammalian species (Genbank accession numbers in table S6). Our taxon choice included more primate representatives than any other order since 4/7 of the viruses infect primate hosts and we wished to evaluate the genes' relationship in this part of the tree in finer detail. All phylogenetic reconstruction was performed in parallelized MrBayes (50, 51). We ran two independent MCMC chains for 10 million generations to ensure convergence (we used the effective sample size of the posterior probability as a proxy for 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 probability below 85 for the nodes relevant to our investigation). The best evolutionary model for each alignment was determined using JModelTest 2 (52) or ProtTest 3 (53) according to the corrected Akaike information criterion. The tree shown in figure 2 was reconstructed from a 266 amino acid alignment of MHC class 1 genes. The CrHV2 homolog was derived from the 3 separate predicted ORFs R1, R2 and R3 (RV138-140 in table S3). In figure 3,

the IL17 tree was reconstructed using a 144 amino acid alignment of 36 mammalian IL17 genes and a homolog in SaHV2 (RV134 in table S3). The DHFR tree in figure 6A was reconstructed from a 176 amino acid alignment of DHFR for 35 mammalian species and homologs from SaHV2, MaHV5 and HHV8 (RV10 in table S3). The CD59 tree in figure 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 figure 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 figure 5C was constructed using a 93 nucleotide alignment of 35 mammalian CCL3 genes, an MaHV5 homolog and 3 homologs identified in HHV8 (RV13, RV14, RV17 in table S3). The CASP8 tree shown in figure S1 was reconstructed from a nucleotide alignment of 36 mammalian genes and their homologs in 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. low posterior probabilities across the tree and/or no convergence of MCMC chains). We therefore implemented the SRD06 model (separate HKY+G for codon 1+2 and 3) in the case of the IL-6 tree (Fig 4A), CD200 tree (Fig 7A), R1 and Acyl-CoA trees (Fig S1). SRD06 is thought to be more biologically realistic and has been shown to frequently outperform other approaches (54).

## Selection Analysis

We assessed the effect of viral capture on selection pressure 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 (49, 55, 48, 56, 47). We tested for individual amino acid sites under positive selection in the alignment for eight genes (excluding those where 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

562 the tree we constructed for the analysis as input to PAML (a software for the analysis of selection  
563 under a maximum likelihood framework (57)). At least one positively selected site was identified in  
564 all but 2 of the genes tested (table S1). For each of the 8 genes, we tested whether our results  
565 may have been biased by the fact that we chose to include many primate sequences that are very  
566 similar, which could have obscured positively evolving sites (table S5). We therefore repeated the  
567 analysis with only five primate sequences and found that this had a small influence on the results,  
568 where fewer sites were detected in CD59, IL-6 and MHC class II (table S5). We also implemented  
569 the 'branch-site test' to detected positively evolving sites along the viral branch (or internal viral  
570 branch in the trees with multiple viral tips). The goal was to identify the sites that underwent rapid  
571 evolutionary change immediately after being captured, but the analysis revealed no such sites. It  
572 should be noted that the sensitivity and power of this test is highly dependent on the nature of the  
573 dataset, requiring ideal levels of divergence, alignment length and sample size (58). In our case,  
574 we speculate that the size of the data set may have influenced the negative result, especially since  
575 the changes we are trying to detect will have been fleeting in evolutionary terms. It would be  
576 interesting for future studies to simulate the range of data sizes and types necessary to detect  
577 such an event.

578 We used the branch model test implemented in CODEML, where an *a priori* model of selection is  
579 compared against a null hypothesis. For each gene, we tested the null hypothesis that dN/dS is the  
580 same across the tree (table S1; model= $\omega_0$ ), and an alternative model where the viral branch  
581 dN/dS is estimated separately. For three of the genes, there are more than one 'viral branch' for  
582 genes that are found in multiple viruses (homologs of CD200 & CCL3 and DHFR). We therefore  
583 tested three alternative models, allocating either separate dN/dS for the branch internal to the viral  
584 group, estimating a separate dN/dS for all viral branches or designating separate values for the  
585 internal branch and viral tips (table S1). In order to examine the selection dynamics along each  
586 gene in finer detail, we developed and implemented a sliding-window modification of the branch  
587 test. In our modification, rather than comparing the overall dN/dS values for the branches under  
588 consideration, we analyse subsections of the alignment, plotting the dN/dS values along with  
589 whether or not a significant difference is exhibited. A custom script was written to run multiple

590 instances of CODEML for a window of 120 base pairs with a step size of 3 both the null and  
591 alternative hypotheses were run for each window, and a likelihood ratio test was used to determine  
592 significance according to a chi2 test with a p-value threshold of 0.05. This p-value threshold is  
593 being used as a proxy for the magnitude of the effect, meant to highlight the windows that exhibit  
594 the largest differences in dN/dS. They should not be interpreted as an indicator of significance,  
595 except in the case of the overall test that identifies the general trend. Rather, the sliding window  
596 approach is an exploratory tool that illuminates heterogeneity of the effect we are detecting.  
597 Window sizes of 30, 60, 90 and 150 were also attempted but these resulted in poorer results –  
598 shorter windows produced fewer stretches of statistically significant differences in dN/dS and a  
599 larger window resulted in lower resolution for shorter genes. Note that since the same alignment is  
600 used for the standard analysis and the sliding window modification, the x-axis scale in all charts  
601 shown in fig. 2, 3, 4, 6 and 7 is the same. The scale is also the same along the y-axis, and the tick  
602 spacing is optimized for readability in each case.”

### 603 **Sequence composition analysis, structural modeling & co-phylogeny analysis**

604 For each alignment in our main analysis, we measured the gene lengths, CpG bias, all four  
605 mononucleotide and 16 dinucleotide frequencies using a custom script. These variables were used  
606 in a principal components analysis to identify whether a pattern could be detected that  
607 distinguished viral genes from their host homologs. We examined the first three principal  
608 components in all genes analysed (Fig 9), which in each case collectively account for >90% of the  
609 variance. In addition to the viral gene under investigation and its homologs in host species, we also  
610 included the rest of the genes in the viral genome to which the gene belongs to assess their  
611 location in the plot relative to the host gene and viral homolog.

612 We used the Phyer2 web portal to model the tertiary structure of viral homologs to evaluate the  
613 potential functional significance of any sequence divergence since HGT (59). To enable direct  
614 comparison with the results of the selection analyses, we used the same amino acid sequence as  
615 the PAML input (i.e. with some trimmed regions that could not be aligned to host homologs). We  
616 generated an overlap of the model generated with the database template that is most likely a  
617 homolog of the query in the UCSF Chimera package (60) (Fig 2-5). 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 colour spectrum of blue to red in where each increments represents an increase of dN/dS of 0.25, where the darkest blue is equivalent to 0 and the darkest red is 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 is to a reference structure. For this reason, we do not consider the models as 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 co-phylogeny analysis using JANE 4.0, which is an event-based method of reconciling tree topologies of hosts and parasites (61). The method works by considering five possible evolutionary events: co-speciation, duplication, loss, host switching and failure to diverge, and reconstructs possible evolutionary histories that could explain the data with a minimum cost of events. In our analysis, we used the default cost-scheme of 0,1,2,1,1 for co-speciation, 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 are comparing the congruence of tree topology between the viral phylogeny and a gene tree rather than host-parasite. We calculated the distribution of costs for random sample solutions for 1000 samples, using random tip mapping as the randomization method.

## References

1. Alcami A. 2003. Viral mimicry of cytokines, chemokines and their receptors. *Nat Rev Immunol* 3:36–50.
2. Holzerlandt R, Orenco C, Kellam P, Albà MM. 2002. Identification of new herpesvirus gene homologs in the human genome. *Genome Res* 12:1739–48.



- 646 3. Schönrich G, Abdelaziz MO, Raftery MJ. 2017. Herpesviral capture of immunomodulatory host genes. *Virus*  
647 *Genes* 1–12.
- 648 4. Ouyang P, Rakus K, van Beurden SJ, Westphal AH, Davison AJ, Gatherer D, Vanderplasschen AF. 2014. IL-10  
649 encoded by viruses: a remarkable example of independent acquisition of a cellular gene by viruses and its  
650 subsequent evolution in the viral genome. *J Gen Virol* 95:245–62.
- 651 5. McGeoch DJ, Davison AJ, Dolan A, Gatherer D, Sevilla-Reyes EE. 2008. Molecular Evolution of the  
652 Herpesvirales, p. 447–475. *In* Domingo, E, Holland, JJ (eds.), *Origin and Evolution of Viruses*. Academic Press.
- 653 6. Feschotte C, Gilbert C. 2012. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat*  
654 *Rev Genet* 13:283–296.
- 655 7. Holmes EC. 2011. The Evolution of Endogenous Viral Elements. *Cell Host Microbe* 10:368–377.
- 656 8. Katzourakis A, Gifford RJ. 2010. Endogenous viral elements in animal genomes. *PLoS Genet* 6:e1001191.
- 657 9. Katzourakis A. 2013. Paleovirology: inferring viral evolution from host genome sequence data. *Phil Trans R Soc*  
658 *B* 368:20120493-.
- 659 10. Aswad A, Katzourakis A. 2012. Paleovirology and virally derived immunity. *Trends Ecol Evol* 27:627–636.
- 660 11. Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of  
661 endogenous retroviruses. *Science* (80- ) 351:1083–1087.
- 662 12. Pellett PE, Biggin MD, Barrell B, Roizman B. 1985. Epstein-Barr virus genome may encode a protein showing  
663 significant amino acid and predicted secondary structure homology with glycoprotein B of herpes simplex virus 1.  
664 *J Virol* 56:807–13.
- 665 13. Ehlers B, Dural G, Yasmum N, Lembo T, de Thoisy B, Ryser-Degiorgis M-P, Ulrich RG, McGeoch DJ. 2008.  
666 Novel mammalian herpesviruses and lineages within the Gammaherpesvirinae: cospeciation and interspecies  
667 transfer. *J Virol* 82:3509–3516.
- 668 14. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence  
669 with rearrangements. *Genome Res* 14:1394–403.
- 670 15. Loh J, Zhao G, Nelson CA, Coder P, Droit L, Handley SA, Johnson LS, Vachharajani P, Guzman H, Tesh RB,  
671 Wang D, Fremont DH, Virgin HW. 2011. Identification and sequencing of a novel rodent gammaherpesvirus that  
672 establishes acute and latent infection in laboratory mice. *J Virol* 85:2642–56.
- 673 16. Gu C, Wu L, Li X. 2013. IL-17 family: cytokines, receptors and signaling. *Cytokine* 64:477–85.
- 674 17. Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B, Pontius J, Roelke M,  
675 Rumpler Y, Schneider MPC, Silva A, O'Brien SJ, Pecon-Slattery J. 2011. A molecular phylogeny of living  
676 primates. *PLoS Genet* 7:e1001342.
- 677 18. Opazo JC, Wildman DE, Prychitko T, Johnson RM, Goodman M. 2006. Phylogenetic relationships and  
678 divergence times among New World monkeys (Platyrrhini, Primates). *Mol Phylogenet Evol* 40:274–80.
- 679 19. Kishimoto T. 2010. IL-6: from its discovery to clinical applications. *Int Immunol* 22:347–52.
- 680 20. Markine-Goriaynoff N, Georgin J-P, Goltz M, Zimmermann W, Broll H, Wamwayi HM, Pastoret P-P, Sharp PM,  
681 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* 77:1784–92.
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* 87:1509–19.
22. Berg J, Tymoczko J, Stryer L. 2002. *Biochemistry*, 5th ed. W.H. Freeman & Co Ltd., New York.
23. Cinquina CC, Grogan E, Sun R, Lin SF, Beardsley GP, Miller G. 2000. Dihydrofolate reductase from Kaposi's sarcoma-associated herpesvirus. *Virology* 268:201–17.
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.
25. Kirkitadze MD, Barlow PN. 2001. Structure and flexibility of the multiple domain proteins that regulate complement activation. *Immunol Rev* 180:146–161.
26. Rezaee SAR, Cunningham C, Davison AJ, Blackbourn DJ. 2006. Kaposi's sarcoma-associated herpesvirus immune modulation: an overview. *J Gen Virol* 87:1781–804.
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.
28. Zong JC, Ciufo DM, Alcendor DJ, Wan X, Nicholas J, Browning PJ, Rady PL, Tying 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* 73:4156–70.
29. Sanjuán R. 2012. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. *PLoS Pathog* 8:e1002685.
30. Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* 9:267–76.
31. Katzourakis A, Aswad A. 2016. Evolution: Endogenous Viruses Provide Shortcuts in Antiviral Immunity. *Curr Biol* 26:R427–R429.
32. Holmes EC. 2011. What does virus evolution tell us about virus origins? *J Virol* 85:5247–51.
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.
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.
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* 150:831–41.
36. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet* 8:857–68.

- 718 37. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D,  
719 Lander ES. 2006. Positive natural selection in the human lineage. *Science* 312:1614–20.
- 720 38. Elde NC, Malik HS. 2009. The evolutionary conundrum of pathogen mimicry. *Nat Rev Microbiol* 7:787–797.
- 721 39. Schmid K, Yang Z. 2008. The trouble with sliding windows and the selective pressure in BRCA1. *PLoS One*  
722 3:e3746.
- 723 40. Bernatchez L, Landry C. 2003. MHC studies in nonmodel vertebrates: what have we learned about natural  
724 selection in 15 years? *J Evol Biol* 16:363–377.
- 725 41. Spurgin LG, Richardson DS. 2010. How pathogens drive genetic diversity: MHC, mechanisms and  
726 misunderstandings. *Proc Biol Sci* 277:979–88.
- 727 42. Babić M, Krmpotić A, Jonjić S. 2011. All is fair in virus-host interactions: NK cells and cytomegalovirus. *Trends*  
728 *Mol Med* 17:677–85.
- 729 43. Pyzik M, Dumaine A, Dumaine AA, Charbonneau B, Fodil-Cornu N, Jonjic S, Vidal SM. 2014. Viral MHC class I-  
730 like molecule allows evasion of NK cell effector responses in vivo. *J Immunol* 193:6061–9.
- 731 44. Holzerlandt R, Orengo C, Kellam P, Albà MM. 2002. Identification of new herpesvirus gene homologs in the  
732 human genome. *Genome Res* 12:1739–48.
- 733 45. Lambris JD, Ricklin D, Geisbrecht B V. 2008. Complement evasion by human pathogens. *Nat Rev Microbiol*  
734 6:132–42.
- 735 46. Nicholas J. 2005. Review: Human Gammaherpesvirus Cytokines and Chemokine Receptors. *J Interf Cytokine*  
736 *Res* 25:373–383.
- 737 47. Yang Z, Zhang Y, Wafula EK, Honaas LA, Ralph PE, Jones S, Clarke CR, Liu S, Su C, Zhang H, Altman NS,  
738 Schuster SC, Timko MP, Yoder JI, Westwood JH, dePamphilis CW. 2016. Horizontal gene transfer is more  
739 frequent with increased heterotrophy and contributes to parasite adaptation. *Proc Natl Acad Sci* 113:E7010–  
740 E7019.
- 741 48. Treangen TJ, Rocha EPC. 2011. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families  
742 in Prokaryotes. *PLoS Genet* 7:e1001284.
- 743 49. Jayawardane G, Russell GC, Thomson J, Deane D, Cox H, Gatherer D, Ackermann M, Haig DM, Stewart JP.  
744 2008. A captured viral interleukin 10 gene with cellular exon structure. *J Gen Virol* 89:2447–2455.
- 745 50. Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models.  
746 *Bioinformatics* 19:1572–1574.
- 747 51. Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. 2004. Parallel Metropolis coupled Markov chain Monte  
748 Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20:407–15.
- 749 52. Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel  
750 computing. *Nat Methods* 9:772.
- 751 53. Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein  
752 evolution. *Bioinformatics* 27:1164–5.
- 753 54. 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.
55. 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.
56. 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.
57. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–91.
58. Yang Z, dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol* 28:1217–28.
59. 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.
60. 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* 25:1605–12.
61. Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R. 2010. Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms Mol Biol* 5:16.

## Figure legends

**Figure 1. Whole genome alignment of rhadinoviruses.** The figure depicts a scaled cartoon of an alignment of seven rhadinoviral genomes, where boxes represent 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 grey or white in alternating fashion. The variable regions are further exploded for clarity, highlighting the genes that were analysed 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 colour and annotated in the zoomed-in panels below the alignment. Virus abbreviations are: HHV8: *Human herpesvirus 8* (also known as *Kaposi's sarcoma herpesvirus* or KSHV), MaHV5: *Macacine gammaherpesvirus 5*, BoHV4: *Bovine gammaherpesvirus 4*, SaHV2: *Saimirine gammaherpesvirus 2*, AtHV3: *Ateline gammaherpesvirus 3*, MuHV4: *Murid gammaherpesvirus 4*, CrHV2: *Cricetid gammaherpesvirus 2* (also known as *Rodent herpesvirus Peru* or RHVP)

785

786 **Figure 2. Capture of a Major histocompatibility complex (MHC) gene by CrHV2.** Panel A:

787 Bayesian phylogenetic reconstruction of mammalian MHC class I genes with the CrHV2 homolog.

788 The CrHV2 branch length has been truncated to fit in the figure and the total branch length is

789 indicated in brackets. The scale bar represents 0.1 substitutions per site. The numbers at each

790 node represent posterior probability and values below 80 are not shown. This tree shows that the

791 CrHV2 homolog and the genes of *M. musculus* and *R. norvegicus* group together with 100

792 probability. Panel B: The graph depicts a collation of a number of selection analysis results. The

793 positions of known domains are labelled above the graph. The grey line indicates the values of

794 dN/dS obtained for each residue. Sites shown to be under significant positive selection according

795 to the Bayes empirical Bayes analysis (a method to calculate the posterior probability that a site

796 belongs to a dN/dS value category using a 95% threshold) are indicated by a grey dot along the

797 line. The green (host) and blue (virus) lines represent the results of the sliding window branch test

798 analysis. The thick line indicates where the branch test rejected the null hypothesis of a single

799 dN/dS according to a likelihood ratio test. Panel C: Structural model of the MHC homolog (gold)

800 with the best matching template (mouse MHC) in the database (blue). There is a helix at both

801 termini that could not be modelled suggesting that they could have been lost in the viral copy (the

802 corresponding region is not homologous). This suggests that the viral copy does not function as a

803 transmembrane protein since the C-terminus helix is needed for this. Panel D: The same structural

804 model as in panel C, but only showing the antigen binding cleft coloured according to the dN/dS

805 values for each residue. Nearly all of the positively evolving sites are found in this region.

806

807 **Figure 3. Host-derived IL-17 in MaHV5, HHV8 and SaHV2.** Panel A: Bayesian phylogenetic

808 reconstruction of mammalian IL-17 and a homologous sequence in SaHV2. The scale bar

809 represents 0.1 substitutions per site. The numbers at each node represent posterior probability and

810 values below 80 are not shown. The tree shows that the viral gene groups most closely to the new

811 world monkeys *C. jacchus* and *S. boliviensis*, the latter being the virus' current host. Panel B: The

812 graph depicts a collation of a number of selection analysis results. The grey line indicates the

813 values of dN/dS obtained for each residue. No sites under significant positive selection according

814 to the Bayes empirical Bayes analysis were identified. The green (host) and blue (virus) lines  
815 represent the results of the sliding window branch test analysis. The thick line indicates where the  
816 branch test rejected the null hypothesis of a single dN/dS according to a likelihood ratio test. The  
817 conserved domains and other regions in IL-17 are annotated in brown. Panel C: Structural model  
818 of the IL-17 homolog (gold) with the best matching template (human IL17) in the database (blue),  
819 showing precise overlap for most of the structure. Panel D: The same structural model as in panel  
820 C without the template model, and coloured according to the dN/dS values for each residue.

821

822 **Figure 4. Host-derived IL-6 in MaHV5, HHV8 and SaHV2.** Panel A: Bayesian phylogenetic  
823 reconstruction of mammalian IL-6 and homologous sequences in HHV8 and MaHV5. The scale  
824 bar represents 0.1 substitutions per site. The numbers at each node represent posterior probability  
825 and values below 80 are not shown. The tree shows that the viral genes group most closely to the  
826 tree shrew (*T. chinensis*) with 99 probability despite belonging to viruses that infect primates. This  
827 could either indicate that a cross species transfer had occurred in the history of the virus, or that  
828 long-branch attraction has resulted in this unusual topology. Panel B: The graph depicts a collation  
829 of a number of selection analysis results, but these did not include the MaHV5 gene since it could  
830 not be aligned for the entire length (219 gaps in a 384 nucleotide alignment). The grey line  
831 indicates the values of dN/dS obtained for each residue. Sites shown to be under significant  
832 positive selection according to the bayes empirical bayes analysis are indicated by a grey dot  
833 along the line. The green (host) and blue (virus) lines represent the results of the sliding window  
834 branch test analysis. The thick line indicates where the branch test rejected the null hypothesis of a  
835 single dN/dS according to a likelihood ratio test. Panel C: Structural model of both IL-6 homologs  
836 (gold) with the best matching template (mouse IL-6) in the database (blue). Note that as well as  
837 sequence divergence, the MaHV5 structure is also more divergent than the HHV8 counterpart.  
838 Panel D: The structural model of HHV8 as in panel C without the template model, and coloured  
839 according to the dN/dS values for each residue. Residues with a dN/dS above 1 are distributed  
840 throughout the structure with no discernable pattern.

841

842 **Figure 5. Phylogenetics & structural models of captured genes in HHV8, BoHV4 & SaHV2.**

843 This figure contains the genes from our main analysis for which a sliding window analysis was not  
844 conducted because of a short alignment length in the case of CD59 and CCL3 and because the  
845 overall branch test was not significant for the C2GnT-M. Panel A: Bayesian phylogenetic  
846 reconstruction of mammalian CD59 glycoprotein genes and homologous sequences in SaHV2.  
847 The scale bar represents 0.1 substitutions per site. The numbers at each node represent posterior  
848 probability and values below 80 are not shown. The tree shows that the viral gene groups with *S.*  
849 *boliviensis* with 97 probability, indicating that the gene is most likely derived from this monkey,  
850 which is the virus' current host. The relatively short branch length is also an indication that the  
851 capture has occurred recently. Panel B: Bayesian phylogenetic reconstruction of mammalian  
852 C2GnT-M genes with a homolog from BoHV4. The scale bar represents 0.1 substitutions per site.  
853 The numbers at each node represent posterior probability and values below 80 are not shown.  
854 This tree confirms previous studies that have reconstructed trees with comparable topology. This  
855 has been previously interpreted as evidence that the origin of the gene is *S. caffer*. . Panel C:  
856 Bayesian phylogenetic reconstruction of mammalian CCL3 genes with homologous genes from  
857 MaHV5 and three copies in HHV8. The scale bar represents 0.1 substitutions per site. The  
858 numbers at each node represent posterior probability and values below 80 are not shown. Panels  
859 D, E & F: Structural models of CD59, C2GnT-M and CCL3 are represented in gold. The best  
860 matching database templates for CD59 and C2GnT-M were xx and mouse C2GnT-L. In the case  
861 of the CCL3 homologues, the best matching template was the human thymus and activation-  
862 regulated chemokine (TARC).

863  
864 **Figure 6. Host-derived DHFR in SaHV2, HHV8 & MaHV5.** Panel A: Bayesian phylogenetic

865 reconstruction of mammalian DHFR and homologous sequences in SaHV2, HHV8 and MaHV5.  
866 The scale bar represents 0.1 substitutions per site. The numbers at each node represent posterior  
867 probability and values below 80 are not shown. The tree shows that the viral gene groups within  
868 primates with 91 probability and is probably most closely related to great ape genes as there is a  
869 posterior probability of 96 that the virus and the clade of great apes are sister clades. Panel B: The  
870 graph depicts a collation of a number of selection analysis results. The grey line indicates the



871 values of dN/dS obtained for each residue. No sites under significant positive selection according  
872 to the bayes empirical bayes (BEB) analysis were identified. The green (host), blue (virus) and  
873 orange (viral stem branch) lines represent the results of the sliding window branch test analysis.  
874 The thick line indicates where the branch test rejected the null hypothesis of a single dN/dS  
875 according to a likelihood ratio test. dN/dS estimates of infinity for the viral stem branch are not  
876 shown for clarity. Panel C: Structural model of the DHFR homolog (gold) with the best matching  
877 template (human dihydrofolate reductase) in the database (blue), showing precise overlap for the  
878 entire structure. Panel D: The same structural model as in panel C without the template, and  
879 coloured according to the dN/dS values for each residue.

880

881 **Figure 7. Host-derived CD200 in SaHV2, HHV8 & MaHV5.** Panel A: Bayesian phylogenetic  
882 reconstruction of mammalian CD200 genes and homologous sequences in HHV8 and MaHV5.  
883 The scale bar represents 0.1 substitutions per site. The numbers at each node represent posterior  
884 probability and values below 80 are not shown. The viral genes are most closely related to new  
885 world monkey host genes from *C. jacchus* and *S. boliviensis* with 92 probability, and more closely  
886 related to *C. jacchus* with 82 probability. Panel B: The graph depicts a collation of a number of  
887 selection analysis results. The grey line indicates the values of dN/dS obtained for each residue.  
888 One site under significant positive selection according to the BEB analysis is indicated by a grey  
889 dot. The green (host), blue (virus) and orange (viral stem branch) lines represent the results of the  
890 sliding window branch test analysis. The thick line indicates where the branch test rejected the null  
891 hypothesis of a single dN/dS according to a likelihood ratio test. dN/dS estimates of infinity for the  
892 viral stem branch are not shown for clarity. Panel C: Structural model of HHV8 (upper) and MaHV5  
893 (lower) homologs (gold). The best matching template in the database was mouse CD200 (blue).  
894 Panel I: The structural model of HHV8 and MaHV5 as in panel D without the template model, and  
895 coloured according to the dN/dS values for each residue.

896

897 **Figure 8. Co-phylogeny analysis for FGAM synthase and CCP.** Panel A: The results of the co-  
898 phylogeny analysis for FGAM synthase homologs in rhadinoviruses. The tree shown is a maximum  
899 likelihood phylogeny reconstructed from a concatenation of the 6 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 duplications, losses and rearrangements have occurred in the clade. The viral tree is shown with reconstructed events superimposed at each node. The sketches represent a simplified viral genome with FGAM 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. Panel B: Phylogenetic tree of all FGAM genes from rhadinoviruses (including copies). Dark taxon labels indicate 3' end loci. Scale bar represents substitutions per site. Panel C: A histogram of the cost of a random sample solution, showing that the co-phylogeny reconstruction for FGAM at a cost of 8 compared to the next best random cost at 27 ( $p < 0.0001$ ). We implemented a cost scheme of 0,1,2,1,1 for co-speciation, duplication, loss, host switching and failure to diverge, respectively. Panel D & E: Results of the co-phylogeny analysis for CCP homologs superimposed on the virus tree as a cladogram (the branch lengths of coloured dotted lines is 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) solution is shown with (Panel E) and without (Panel D) horizontal gene transfer allowed. The cost for the CCP solutions with and without HGT allowed lies within the distribution of costs for random samples ( $p = 0.151$  and  $p = 0.029$ , respectively). The same cost scheme as in panel C is used. Virus abbreviations are: CalHV3: *Calitrichine herpesvirus 3*, HHV4: *Human herpesvirus 4* (also known as *Epstein-Barr virus* or *EBV*), HHV8: *Human herpesvirus 8*, MaHV5: *Macacine gammaherpesvirus 5*, BoHV4/6: *Bovine gammaherpesvirus 4/6*, SaHV2: *Saimirine gammaherpesvirus 2*, AtHV3: *Ateline gammaherpesvirus 3*, MuHV4: *Murid gammaherpesvirus 4*, CrHV2: *Cricetid gammaherpesvirus 2*, EqHV2: *Equine herpesvirus 2*, WmHV: *Wood mouse herpesvirus*, PoHV2/3: *Porcine Herpesvirus 2/3*.

**Figure 9. Principal components analysis of sequence composition, length and CpG bias.**

Principal components analysis of the 8 main genes investigated in this study. The variables used were all four mononucleotide and sixteen dinucleotide frequencies, in addition to CpG bias and gene length. For each gene, the scores for the first 3 principal components (accounting for most of

929 the variance) are plotted for the mammalian host genes as small solid dark green points. All the  
930 viral genes in the genome are shown as either pink, blue or lime green, highlighting the host-  
931 derived gene in question as a larger, dark point of the same colour. In the case of Dihydrofolate  
932 reductase, CD200 and CCL3, host derived homologs are observed in multiple rhadinoviruses, and  
933 so each virus is assigned a different colour. In all of the genes except for C2GnT-M, there is a  
934 clear distinction in variable space between the viral homologs and the host genes from which they  
935 originate. See table S4 for a detailed breakdown of the variable loadings for each principal  
936 component.

937

938





















