On phylogenetic relationships among major lineages of the Gammaherpesvirinae

Duncan J. McGeoch, Derek Gatherer and Aidan Dolan

Medical Research Council Virology Unit, Institute of Virology, University of Glasgow, Church Street, Glasgow G11 5JR, UK

Phylogenetic relationships within the subfamily Gammaherpesvirinae of the family Herpesviridae were investigated for three species in the genus Lymphocryptovirus (or γ1 group) and nine in the genus Rhadinovirus (or γ2 group). Alignments of amino acid sequences from up to 28 genes were used to derive trees by maximum-likelihood and Bayesian Monte Carlo Markov chain methods. Two problem areas were identified involving an unresolvable multifurcation for a clade within the γ2 group, and a high divergence for Murid herpesvirus 4 (MHV4). A robust final tree was obtained, which was valid for genes from across the virus genomes and was rooted by reference to previous analyses of the whole family Herpesviridae. This tree comprised four major lineages: the γ1 group of primate viruses; a clade of artiodactyl γ2 viruses; a clade of perissodactyl γ2 viruses; and a clade of γ2 viruses with a multifurcation at its base and containing Old World and New World primate viruses, Bovine herpesvirus 4 and MHV4. Developing previous work it was proposed, on the basis of similarities between the gammaherpesvirus tree and the tree of corresponding mammalian hosts, that the first three of these major viral lineages arose in a coevolutionary manner with host lineages, while the fourth had its origin in an ancient interspecies transfer. Transfer of dates from mammalian palaeontology then allowed estimation of dates for nodes in the gammaherpesvirus tree.

INTRODUCTION

Understanding of evolutionary relationships among members of the family Herpesviridae has improved greatly over the last decade, with application of computer-based methods of phylogenetic analysis to the increasing amounts of DNA sequences available for herpesvirus genomes (McGeoch & Cook, 1994; McGeoch et al., 1995, 2000). Mammalian herpesviruses are now clearly seen all to have a common evolutionary origin, from which they have diverged widely to form three subfamilies, the Alpha-, Beta- and Gammaherpesvirinae (Minson et al., 2000). Herpesviruses with avian and reptilian hosts so far characterized belong to the Alphaherpesvirinae (Minson et al., 2000; McGeoch & Gatherer, 2005), but herpesviruses of amphibians and fish have only a very distant relationship to the three subfamilies, and the one characterized invertebrate herpesvirus (of oysters) makes a third distinct grouping (Davison, 2002). Within subfamilies of the Herpesviridae a substantial subset of the phylogenetic tree’s branching features reflect patterns in the tree for lineages of mammalian hosts, suggesting coevolution of host and virus lineages, and this correspondence has enabled a timescale to be constructed for herpesvirus phylogeny (McGeoch & Cook, 1994; McGeoch et al., 1995; McGeoch & Gatherer, 2005).

The subfamily Gammaherpesvirinae is presently divided into two genera, namely Lymphocryptovirus (or γ1 group) and Rhadinovirus (or γ2 group). The γ1 group contains Epstein–Barr virus (EBV) (see Table 1 for abbreviations of virus names) and its primate relatives, while the γ2 group contains herpesviruses with hosts of many mammalian taxa; those treated in this paper have primate, rodent, and artiodactyl and perissodactyl ungulate hosts; human herpesvirus 8 (HHV8) is the single known human virus in the γ2 group. Recent years have seen much activity in discovery and characterization of gammaherpesviruses. Until recently only Old World primate (OWP) viruses were known in the γ1 group, but a γ1 virus of a New World primate (NWP) has now been described (Callitrichine herpesvirus 3, CHV-3) (Rivailler et al., 2002a). In the γ2 group, many new viruses have been detected recently, with hosts that include primate, ungulate and carnivore species (for instance: Rovnak et al., 1998; Greensill et al., 2000; Lacoste et al., 2000; Schultz et al., 2000; Banks et al., 2002; Kleiboeker et al., 2002; Ehlers & Lowden, 2004).

Previous work on phylogeny of the Gammaherpesvirinae fell short of achieving a tree that was completely resolved and satisfactorily interpreted in terms of possible virus–host coevolution (McGeoch et al., 2000; McGeoch, 2001); complications included aberrant behaviour of one virus (Murid herpesvirus 4, MHV4) apparently associated with a particularly high level of sequence change, and inability
to completely resolve branching order within the γ2 group. Our aim for this paper was to reanalyse gammaherpesvirus phylogeny with increased discrimination, by treating only gammaherpesvirus sequences and thereby enabling use of a larger set of genes than was possible in our earlier analyses, which encompassed all three subfamilies. We succeeded in constructing a robust tree (although still with a multifurcation) and were able to propose a global interpretation of the relationship of gammaherpesvirus evolution to that of host lineages.

**METHODS**

**General computational handling of molecular sequences.** Genomic sequences were obtained from the EMBL library, as listed in Table 1. Sequence handling used the GCG (Accelrys) and EMBOSS packages. Alignment of amino acid sequences was carried out by combined use of three alignment programs, as described previously (McGeoch et al., 2000). Positions in alignments with gaps introduced by the alignment programs were removed before use for phylogenetic inference, as were any regions too diverged to align.

**Computational analyses of phylogenetic trees.** The PHYLIP package (version 3.6; Felsenstein, 1989) was used to construct trees by the neighbour-joining method and to display and manipulate trees. PROTMIL in the MOLPHY package (Adachi & Hasegawa, 1994) was used in preliminary maximum-likelihood (ML) evaluations of alignments of amino acid sequences and to generate lists of tree topologies. TREEADDER (McGeoch & Gatherer, 2005) was used to comprehensively interpolate additional species into tree topologies.

For inference of phylogenetic trees, measures of divergence between amino acid sequences were in all cases made using the matrix of Jones et al. (1992). In-depth phylogenetic analyses were undertaken by two approaches. In the first, CODEML in the PAML package (version 3.14; Yang, 1997) was used to make ML evaluations of aligned sets of amino acid sequences for sets of candidate trees, mostly with a discrete gamma distribution of eight classes of substitution rate across sites. Files output by CODEML containing log likelihoods for alignment sites were then processed by programs in the CONSEL package (version 0.1f; Shimodaira, 2002) to make ML evaluations of aligned sets of amino acid sequences for sets of candidate trees, mostly with a discrete gamma distribution of eight classes of substitution rate across sites. Files output by CODEML containing log likelihoods for alignment sites were then processed by programs in the CONSEL package (version 0.1f; Shimodaira, 2002) to score trees by the approximately unbiased (AU) test and by multi-scaled bootstrap proportions (BP) (Shimodaira, 2002).

In the second approach, Bayesian analysis using Monte Carlo Markov chains (BMCMC) was carried out on amino acid sequence alignments with MrBayes (version 3; Ronquist & Huelsenbeck, 2003), to generate posterior probability (PP) distributions of trees. Starting trees were randomly chosen and multiple runs of the program were generally made with different starting trees, to check convergence of the process. The program’s defaults for prior probability settings were used. BMCMC processes incorporated a discrete gamma distribution of four classes of substitution rate across sites, included one ‘cold’ and three ‘heated’ chains, and were run for 250 000 or 500 000 generations. Output trees were sampled every 50 or 100 generations and typically the first 1000 trees collected were discarded to allow the process to reach stationarity.

**Estimations of dates for phylogenetic events.** Dates for nodes in herpesvirus trees were derived on the basis of equating particular nodes to palaeontological dates in host lineages. Two methods were then employed. The first was to calculate by CODEML a molecular clock version of the rooted ML tree. This enforces a single rate of change across all branches, and the rate was then converted to calendar time by regression of the divergences for calibrating nodes against the calibration dates (using MINITAB). The second method used the program r8s (version 1.5; Sanderson, 2003), which aims to smooth differences in rates for branches of previously computed
trees, without imposing the globally uniform rate of a molecular clock. The penalized likelihood and quasi-newtonian optimization options of the program were used.

**RESULTS**

**Identification of gammaherpesvirus gene sets for phylogenetic analysis**

Complete or largely complete genome sequences were available for 12 species of the Gammaherpesvirinae, as listed in Table 1. Three species are in the γ1 group, namely CHV3, EBV and Rhesus lymphocryptovirus (RLV), and the other nine are in the γ2 group. The amino acid sequences from 50 orthologous sets of genes were evaluated for use in phylogenetic analysis. For each gene set an alignment of amino acid sequences was constructed, as described previously (McGeoch et al., 2000). Twenty-eight of the alignments were judged to be of potentially usable quality for inferring phylogenetic trees, by criteria of showing a substantive level of identical residue types at aligned loci and a low incidence of gapping introduced by the alignment procedure, while the remainder were discarded as too diverged to yield reliable trees. Regions of low quality alignment and loci with a gap character in any sequence were then removed from each of the 28 alignment sets. Characteristics of the 28 sets are listed in Table 2. Twenty-one sets had all 12 species represented and seven had 11 species but lacked a Porcine lymphotropic herpesvirus 1 (PLHV1) member. Phylogenetic analyses were carried out on concatenated subsets of these alignments as described in the following sections.

**Phylogenetic evaluation of an eight-gene alignment of amino acid sequences**

In previous work we identified eight genes that had orthologues in all sequenced genomes of mammalian and avian herpesviruses of the *Alpha*- and *Beta* - and *Gammaherpesvirinae* and whose encoded amino acid sequences were sufficiently conserved to allow alignments of quality suitable for phylogenetic analysis across the three subfamilies (McGeoch et al., 1995, 2000). In the present study we first examined this primary set for the 12 gammaherpesviruses, comprising genes 06, 07, 08, 09, 25, 29, 44 and 46 in the HVS nomenclature used in Table 2. The concatenated alignment for these eight sets of amino acid sequences, referred to as the $8 \times 12$ set, was 5862 residues long and so is a large dataset by usual standards. Our initial investigation employed the strategy of using increasingly compute-intensive programs to evaluate successively smaller sets of candidate trees, finally employing ML evaluation with allowance for variation of rates across sites, as described previously (McGeoch et al., 2000). From this analysis (results not shown) it became evident that no single tree was clearly best, and that we should instead consider sets of top-scoring trees. We therefore revised our approach to encompass tree scoring methods that represented best current practice (although at high computational cost). The results shown in this paper are primarily based on two types of analysis: first, extensive ML evaluation with a limited number of operational taxonomic units (OTUs) and application of a sophisticated scoring scheme; and second, the BmcmC method. The results obtained initially and by the two subsequent methods were all closely equivalent. From considerations of tree scoring and analytical thoroughness we wished to examine sets of possible trees that were as complete as practicable. However, for feasibility in ML computations we could process exhaustively tree sets based on at most seven distinct OTUs. Virus species were

---

**Table 2. Gammaherpesvirus genes used for phylogenetic analyses**

<table>
<thead>
<tr>
<th>Gene name*</th>
<th>Genomic locus†</th>
<th>Alignment length‡</th>
<th>Group§</th>
</tr>
</thead>
<tbody>
<tr>
<td>06</td>
<td>12584</td>
<td>990</td>
<td>10K</td>
</tr>
<tr>
<td>07</td>
<td>15977</td>
<td>548</td>
<td>10K</td>
</tr>
<tr>
<td>08</td>
<td>18006</td>
<td>636</td>
<td>10K</td>
</tr>
<tr>
<td>09</td>
<td>20507</td>
<td>909</td>
<td>20K</td>
</tr>
<tr>
<td>17</td>
<td>30837</td>
<td>219</td>
<td>30K</td>
</tr>
<tr>
<td>19</td>
<td>33024</td>
<td>420</td>
<td>30K</td>
</tr>
<tr>
<td>21</td>
<td>35195</td>
<td>283</td>
<td>30K</td>
</tr>
<tr>
<td>22</td>
<td>36775</td>
<td>396</td>
<td>30K</td>
</tr>
<tr>
<td>25</td>
<td>41893</td>
<td>1280</td>
<td>40K</td>
</tr>
<tr>
<td>26</td>
<td>46024</td>
<td>292</td>
<td>40K</td>
</tr>
<tr>
<td>29 exon 2</td>
<td>48089</td>
<td>349</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>49443</td>
<td>150</td>
<td>40K</td>
</tr>
<tr>
<td>33</td>
<td>51333</td>
<td>274</td>
<td>50K</td>
</tr>
<tr>
<td>29 exon 1</td>
<td>52315</td>
<td>198</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>53222</td>
<td>257</td>
<td>50K</td>
</tr>
<tr>
<td>36</td>
<td>54499</td>
<td>301</td>
<td>50K</td>
</tr>
<tr>
<td>37</td>
<td>55794</td>
<td>430</td>
<td>50K</td>
</tr>
<tr>
<td>39</td>
<td>57425</td>
<td>323</td>
<td>50K</td>
</tr>
<tr>
<td>42</td>
<td>60572</td>
<td>210</td>
<td>60K</td>
</tr>
<tr>
<td>43</td>
<td>61335</td>
<td>503</td>
<td>60K</td>
</tr>
<tr>
<td>44</td>
<td>63001</td>
<td>714</td>
<td>60K</td>
</tr>
<tr>
<td>46</td>
<td>66168</td>
<td>238</td>
<td>60K</td>
</tr>
<tr>
<td>54ll</td>
<td>74206</td>
<td>214</td>
<td>70K</td>
</tr>
<tr>
<td>55</td>
<td>75104</td>
<td>135</td>
<td>70K</td>
</tr>
<tr>
<td>56</td>
<td>75685</td>
<td>661</td>
<td>70K</td>
</tr>
<tr>
<td>57</td>
<td>78395</td>
<td>241</td>
<td>70K</td>
</tr>
<tr>
<td>60</td>
<td>82273</td>
<td>298</td>
<td>80K</td>
</tr>
<tr>
<td>61</td>
<td>83196</td>
<td>661</td>
<td>80K</td>
</tr>
<tr>
<td>62</td>
<td>85499</td>
<td>253</td>
<td>80K</td>
</tr>
</tbody>
</table>

*Gene names follow the convention used with HVS (Albrecht et al., 1992).
†Genomic locus is given as the left extremity of each gene’s coding sequence (nucleotides) in the genome of HVS (X64346).
‡Alignment length comprises amino acid residues as used for phylogenetic analysis.
§Group denotes subset of genes in phylogenetic scan across genome; see Fig. 3.
||No PLHV1 sequence was available for genes 54 to 62.
therefore lumped into OTUs in cases that we regarded as uncontentious, based on previous analyses. The OTUs defined are illustrated by Fig. 1(a), which shows a tree for the $8 \times 12$ set derived as a starting point by the simple clustering method of neighbour-joining with bootstrap analysis. For ML analyses, the three $\gamma_1$ viruses (CHV3, EBV and RLV) were assigned to a single OTU with the NWP virus CHV3 as sister group to the two OWP viruses, the $\gamma_2$ OWP HHV8 and Rhesus rhadinovirus (RRV) were assigned to one OTU, as were the NWP herpesvirus atelis (HVA) and herpesvirus saimiri (HVS), and the artiodactyl Alcelaphine herpesvirus 1 (AHV1) and PLHV1. This gave a total of seven OTUs, for which 945 bifurcating unrooted trees are possible. We carried out a retrospective check on the validity of using these OTUs, to be described.

ML evaluation of all 945 trees was carried out with the CODEML program of the PAML package, with a discrete gamma distribution of rates across sites. Output from CODEML was then processed with the CONSEL software to compute for each tree the value for the AU test and the multi-scaled bootstrap proportion. Confidence sets of trees were made for which a test measure (AU or BP) was greater than $0.05$. The AU-derived set contained 16 trees and the BP set contained six trees that were a subset of the AU set; Fig. 1(b)–(g) depict the trees from the BP set. The four trees with highest likelihood values from CODEML also had the highest AU and BP scores. Separately, BMCMC analysis of the $8 \times 12$ alignment with the MrBayes program gave a confidence set (PP $>0.01$) which contained only two trees, corresponding to the top-scoring two trees in the AU and BP sets (Fig. 1b and c). Thus, the AU, BP and PP measures identified consistent sets of trees, but with the AU set the largest and the PP set the smallest.

Inspection of the trees in Fig. 1 (b–g) showed that they can all be regarded as based on a constant 10-species tree (with the topology of that shown in Fig. 2a) but with Bovine herpesvirus 4 (BHV4) and MHV4 each appearing at a variety of loci on this tree. The 10 trees in the AU set that are not included in Fig. 1 also conformed to this description. We therefore adopted, as a working hypothesis to facilitate further analysis, the view that inability to identify a single best tree reflected some property of the BHV4 and MHV4 sequences. BHV4 and MHV4 entries were then removed from the $8 \times 12$ alignment and the resulting $8 \times 10$ dataset evaluated with CODEML/CONSEL and MrBayes. This analysis gave confidence sets containing only the single tree shown in Fig. 2(a), and we now refer to this as the ‘standard’ 10-species tree. At this point we carried out a check on the validity of our use of OTUs containing more than one species in the analysis described above: BHV4 and MHV4 were added back to the standard tree topology independently at every possible locus, to give a
set of 323 trees for 12 species (so that trees with BHV4 and MHV4 within loci corresponding to the previous multiple-species OTUs were now included). This set of trees was then evaluated with CODEML/CONSEL, but no novel trees emerged in the resulting confidence sets, indicating that the use of multiple-species OTUs had been in order.

We next carried out analyses that excluded singly either MHV4 or BHV4. Each such 8 × 11 set contained six OTUs (defined as before) and thus 105 possible bifurcating unrooted trees. ML and BMCMC analyses were carried out as before. In addition, for comparative purposes analyses were run that were equivalent except that a uniform rate was imposed for all sites. Results showing BP values are summarized in Fig. 2(b) and (c). For BHV4 (Fig. 2b), the BP values indicate a weak leaning to BHV4 being a sister group to HHV8/RRV; conservatively, they are to be taken as showing that the branching order for BHV4, HHV8/RRV and HVA/HVS is not resolved. BMCMC modelling (both uniform and variable rate), however, gave strong support for BHV4 as sister to HHV8/RRV (PP = 1.000). For MHV4 (Fig. 2c), there is support for possible MHV4 nodes deeper into the tree than with BHV4. The situation is more complicated with MHV4, since the distribution of BP values differs between the analyses with uniform and variable substitution rate across sites: the more elaborate modelling with rate variation gives higher BP values toward the HHV8/RRV clade. BMCMC gave a similar trend, with PP values for MHV4 as a sister of HHV8/RRV of 0.023 and 0.680 in the uniform and variable rate analyses, respectively.

We interpret these findings as follows. First, variability in the BHV4 locus is now seen to represent a straightforward case of very closely spaced nodes for BHV4, HHV8/RRV and HVA/HVS rather than indicating some property truly specific to BHV4. Second, we note that the results for MHV4 locus vary markedly with the analytical method. Harking back to Fig. 1(a) it can be seen that the simplest method employed, that of the neighbour-joining algorithm, gives the deepest node for MHV4, while the two CODEML analyses (Fig. 2c) push the MHV4 node toward the HHV8/RRV clade, the more elaborate version more so. We take this correlation with analytical refinement to indicate that the MHV4 sequences possess some characteristic relative to those of the other species that requires a superior modelling process for optimal outcome, and we presume that the problem lies in an atypically high rate of substitution in MHV4 sequences as evidenced by the long terminal MHV4 branch seen in all trees. Finally, these two factors (of closely spaced nodes and idiosyncratic data) have overlapping ranges of action, so that together they acted in the analyses of Fig. 1 to obscure phylogenetic inference effectively.

The ML trees shown for the 8 × 12 set in Fig. 1 are, except for the low-scoring tree in part (g), compatible with treating the HHV8/RRV, HVA/HVS, BHV4 and MHV4 lineages as branching in indistinguishable order from a single multifurcation. The ML tree with this multifurcation is shown in Fig. 4(a) (in rooted form, to be described below). We regard this as a conservative, well justified representation of the relationships among γ2 lineages. For further discussion we refer to the clade containing HHV8/RRV, HVA/HVS, BHV4 and MHV4 as the MF (for ‘multifurcated’) clade. We postpone to after analysis of other alignment sets the question of whether the BMCMC analysis, with its apparently more focused confidence sets, might yield a better resolved tree.

**Characteristics of gammaherpesvirus trees based on specific genomic regions**

We were interested to ascertain whether the phylogenetic view derived with the 8-gene set was valid across the virus genomes. It might be, for instance, that the phenomenon of apparent accelerated change in MHV4 sequences is specific to certain genes or that recombination between lineages has occurred. For studying relationships across the virus genomes (which are collinear in their order of common genes), the HVS sequence was chosen as the reference for genomic loci. Alignments of 27 from the 28 sets of potentially usable genes were assigned to eight groups, designated 10K–80K, that were defined by the locations of genes in successive 10 kbp segments of the HVS genome; gene 29, which has two widely separated exons, was omitted. The genes in each group are listed in Table 2. The concatenated alignments for groups 10K through 80K contained, successively, 2174, 909, 1318, 1722, 1585, 1665, 1251 and 1212 aa residues. The 70K and 80K groups lacked data for PLHV1. This sampling process...
represented sequences ranging from nt 12584 to nt 86491 within the 112930 nt unique portion of the HVS genome.

Each alignment was analysed by CODEML/CONSEL (with test sets of 945 trees as above) and by MrBayes. These exercises are summarized in Fig. 3 in the condensed form of majority-rule consensus trees from the BMCMC analyses. With one exception, the trees in Fig. 3 show a similar condition to those in Fig. 1 in that they have the topology of the standard 10-species tree (as defined above) plus a variety of loci for BHV4 and MHV4. The exception, Fig. 3(f), is for the 60K group, and here the locations of nodes for Equid herpesvirus 2 (EHV2) and HVA/HVS/BHV4 are reversed relative to the standard tree; we note, however, that this atypical configuration is associated with relatively low support in the PP distribution, and accordingly consider that the result for the 60K group can be discounted. In addition, all the trees in Fig. 3 show a long terminal branch for MHV4. We interpret these results as indicating that, despite considerable noise, the smaller datasets show general consistency with the 8 \times 12 set both in tree topology and in uniquely high level of substitution for the MHV4 terminal branch, so that these features should be taken as conserved across the genomes.

**Optimal representation of phylogenetic relationships**

The overall consistency in trees obtained from sampling across the genomes gave confidence in the validity of using datasets incorporating alignments from many genes to examine phylogeny. In addition to the 8 \times 12 set described above, we examined sets based on 21 genes for all 12 virus species (the 21 \times 12 set, of 9920 aa residues from all the genes in Table 2 except genes 54–62) and on 28 genes for 11 species (the 28 \times 11 set, lacking PLHV1, of 12383 aa residues from all the genes in Table 2). These were evaluated by ML and BMCMC methods as for the 8 \times 12 set, with closely comparable outcomes (not shown): in each case trees in confidence sets were composed of the standard tree (or its 9-species counterpart for the 28 \times 11 set) with variable loci for BHV4 and MHV4, similar to the results for the 8 \times 12 tree in Fig. 1(b–g). And for both the 21 \times 12 and 28 \times 11 sets the ML result reduced to including the MF clade, as for the 8 \times 12 set. The Bayesian PP confidence sets typically have fewer members than those based on AU or BP figures from ML, and with both the 8 \times 12 and 21 \times 12 sets (but not the 28 \times 11 set) BMCMC gave consensus trees with higher resolution than the quadrifurcated version; Fig. 4(c) and (d) show the relevant parts of the trees, for the species of the MF-clade. However, the additional details are not compatible between the two trees, and while the 8 \times 12 set is the smaller it also contains the genes best conserved across the virus family, so we concluded that these BMCMC analyses were displaying ‘overconfidence’ and should be discounted. We thus regard the quadrifurcated tree as presenting the best attainable depiction of topology.

CODEML was used to compute ML branch lengths for the quadrifurcated topology, for all three datasets. The lengths of corresponding branches in these 21 \times 12 and 28 \times 11 trees were compared to those of the 8 \times 12 tree. In both cases proportions were found to be closely maintained across the trees; relative to the 8 \times 12 tree branch lengths in the 21 \times 12 and 28 \times 11 trees are expanded by overall factors of 1·21 and 1·25, respectively (estimated by regression). Thus the 8 \times 12 set, versions of which have previously been used to examine phylogeny across the Herpesviridae, gives a representation of relationships that is consistent within the Gammaherpesvirinae with that from the largest accessible gene sets.
The trees discussed so far are unrooted. From previous phylogenetic analyses that included data from the Alpha-, Beta- and Gammaherpesvirinae we know that the root for the gammaherpesvirus tree lies between the \( \gamma_1 \) and \( \gamma_2 \) groups (McGeoch et al., 1995, 2000). When we compared the \( 8 \times 12 \) ML quadrifurcated tree (Fig. 4a) with the rooted gammaherpesvirus portion of a ML tree for all three subfamilies based on sequences from seven genes and one partial gene of 20 species (data from McGeoch et al., 2000), we observed a close proportionality between pairs of corresponding branch lengths, as illustrated in Fig. 4(b). This overall conservation in proportions then allowed transfer of the root position to the \( 8 \times 12 \) tree with good precision, as indicated in Fig. 4(a). With the addition of the root locus, Fig. 4(a) gives our current optimal representation of gammaherpesvirus phylogeny.

### Coevolution of host and gammaherpesvirus lineages

As mentioned, mammalian herpesvirus lineages show extensive signs of apparent coevolution with host lineages. Interpretation on this basis has been least satisfactory for the Gammaherpesvirinae, in part because of uncertainties in the gammaherpesvirus tree (McGeoch et al., 2000; McGeoch, 2001). We revisited this topic in light of our new robust assignment of major lineages in the tree, to examine what correlations of branching pattern and branch proportions could be made between the host and virus trees, guided by criteria of parsimony and generality. We have identified a scheme that provides an economical account of major features of the gammaherpesvirus tree in terms of host–virus coevolution, and is also consistent with features in the trees for alpha- and betaherpesviruses.

Fig. 5(a) depicts a tree for the higher taxa of relevant host groups, with timescale [as millions of years before the present (Ma)] based primarily on Springer et al. (2003), and Fig. 5(b) presents the gammaherpesvirus tree in a format intended to emphasize proposed correspondences with the host tree. We interpret the following as being of co-evolutionary origin: the OWP and NWP \( \gamma_1 \) lineage, the ruminant AHV1 and said PLHV1 lineages, and the perissodactyl EHV2 lineage. Each of these is known to be populated, beyond the few species represented in our analyses, with other viruses that have hosts in the same group (Rovnak et al., 1998; Ehlers et al., 1999, 2003; Kleiboeker et al., 2002). The only portion of the tree in Fig. 4(a) not thus accounted for is the MF clade, and we interpret this as having a non-coevolutionary origin, involving transfer between host species. The OWP and NWP \( \gamma_2 \) virus lineages could, however, have arisen in a co-evolutionary mode during subsequent development of the MF clade. Treating the MF clade as non-coevolutionary in origin efficiently rationalises a number of features, namely: the position of its origin in the tree; the existence of two groupings of primate gammaherpesviruses; the occurrence of BHV4 in a separate location from other ruminant viruses; and the non-correspondence between the location of MHV4 and the location in the host tree of the rodent lineage (shown in Fig. 5a as a grey line).

We assigned a timescale to this interpretation of the gammaherpesvirus tree by comparison with palaeontological dates for nodes in the mammalian tree. The following dates were taken from Springer et al. (2003): ruminant/pig divergence, 63.8 Ma (for AHV1/PLHV1 node); artiodactyl/perissodactyl divergence, 82.1 Ma (for divergence of AHV1/PLHV1 from EHV2); and primate/ungulate divergence, 94.0 Ma (for divergence of primate \( \gamma_1 \) viruses from AHV1/PLHV1/EHV2). Two methods were used with these figures to obtain divergence dates for other nodes in the quadrifurcated tree, for both the \( 8 \times 12 \) and \( 21 \times 12 \) datasets.
In the first, a ML molecular clock version of the tree was computed with CODEML, then a substitution rate for the whole tree was calculated by regression using the host dates and applied to dating other nodes. In the second method, the program r8s was used, which takes as input a rooted tree with branch lengths and smooths differences among rates on different branches. Results for both approaches are summarized in Table 3.

**DISCUSSION**

We employed two approaches for inference of phylogeny from alignments of amino acid sequences, both powerful and state of the art but both with limitations. For ML evaluation of trees, it is not feasible to treat exhaustively the set of possible trees for more than a modest number of OTUs. Modern tests for evaluation of ML data have supplanted the earlier practice of ranking by likelihood alone, and we found that for inferring confidence sets of trees the multi-scaled bootstrap was a more useful criterion than the AU test. With BMCMC there are two classes of reservation. First, it has complex theoretical underpinnings and is still a new method for phylogenetic inference, which may require time to reach a stable best practice; see discussion by Felsenstein (2004). Second, there is a developing literature on whether the method’s output PP distributions may present an overconfident view (Suzuki et al., 2002; Alfaro et al., 2003; Douady et al., 2003). It was notable in our analyses that the Bayesian PP confidence sets were typically more focused than those from ML. Thus, apparently high-confidence BMCMC results should be treated with caution – as in our comparison of results for the 8 × 12 and 21 × 12 datasets. Nonetheless, application of two distinct methodologies gave valuable added assurance to our analyses.

At the end of these exhaustive analyses, the gammaherpesvirus tree obtained is completely compatible with that from our previous work (McGeoch et al., 2000; McGeoch, 2001), with the addition of four virus species, with assurance of
robustness and generality, and with improved insight into underlying complications. Disentangling the phylogeny of the γ2 group presented difficulties, and recognizing that the nodes for HHV8/RRV, HVA/HVS and BHV4 are effectively coincident and that there is a special problem with the locus of MHV4 were significant steps toward understanding relationships. The overall phylogeny appears to be constant across the genomes, but trees based on small sets of genes were generally noisy and the larger datasets were required for optimal resolution. With the overall constancy of tree topology and proportions seen with the large datasets, we are confident that the optimal tree as presented in Fig. 4(a) is well founded. We regard the tree as comprising four major clades: the γ1 group, the AHV1/PLHV1 lineage, the EHV2 lineage and the MF clade. There are limited sequence data available for many other gammaherpesviruses, but we know of none that demonstrably falls outside these clades.

It is clear that MHV4 sequences have been accumulating changes atypically fast. This effect is not seen with another rodent herpesvirus, the beta herpesvirus murine cytomegalovirus (McGeoch et al., 2000). The visible relative rate of change for MHV4 is, of course, an average over the time since the MHV4 lineage diverged from its sister groups. We cannot, with available data, study substitution characteristics of MHV4 DNA directly; two genome sequences for MHV4 are available (Table 1), but they are of the same strain and very close to identical (Nash et al., 2001) so their comparison is not useful. It is likely that the effect has applied across the genome: it is not a matter of selected mutation in one or a few loci, such as seen with the K1 gene of HHV8 (McGeoch, 2001). The effect could result from relaxed stringency in genomic replication, from more frequent cycles of replication than with other viruses of the group, or from larger virus populations (with ongoing recombination). Whatever its mechanism, this phenomenon may reflect some significant underlying difference between MHV4’s biology and those of other γ2 herpesviruses.

Our date estimates for the gammaherpesvirus tree based on host–virus coevolution (Table 3) are likely to be more precise for γ2 than for γ1 nodes with the calibration system applied. In the γ1 clade, it has previously been apparent that EBV and RLV may be more closely related than the 23 Ma date expected for cospeciation (Ehlers et al., 2003; Gerner et al., 2004). Conversely, in the γ2 group, HHV8 and RRV are more distant than cospeciation counterparts (Greensill et al., 2000; Schultz et al., 2000). These examples emphasize that the concept of host–herpesvirus coevolution should not be pursued down to the species level. If the coevolution hypothesis is valid for higher taxa and longer timescales, then the estimates indicate a considerable antiquity for development of the MF clade, close to the end of the Cretaceous (65 Ma). We presume that the MF clade originated by transfer of a virus between host species. The clade’s immediate ancestor could have been a virus with a perissodactyl or carnivore host (see location of the carnivore lineage in the host tree, shown in Fig. 5a as a grey line). Involvement of a carnivore gammaherpesvirus ancestor would be consistent with our estimates for the development of the MF clade (Table 3) in relation to that for divergence of perissodactyl and carnivore lines (80–4 Ma; Springer et al., 2003).

ACKNOWLEDGEMENTS

We thank M. Goltz and F. Wang for early sight of data and A. Davison for a critical reading of the manuscript. This work was funded by the UK Medical Research Council.

REFERENCES


