Purifying Selection Can Obscure the Ancient Age of Viral Lineages

Joel O. Wertheim*[,1] and Sergei L. Kosakovsky Pond[2]

[1]Department of Pathology, University of California, San Diego
[2]Department of Medicine, University of California, San Diego

***Corresponding author:** E-mail: jwertheim@ucsd.edu.

Associate editor: Alexei Drummond

Abstract

Statistical methods for molecular dating of viral origins have been used extensively to infer the time of most common recent ancestor for many rapidly evolving pathogens. However, there are a number of cases, in which epidemiological, historical, or genomic evidence suggests much older viral origins than those obtained via molecular dating. We demonstrate how pervasive purifying selection can mask the ancient origins of recently sampled pathogens, in part due to the inability of nucleotide-based substitution models to properly account for complex patterns of spatial and temporal variability in selective pressures. We use codon-based substitution models to infer the length of branches in viral phylogenies; these models produce estimates that are often considerably longer than those obtained with traditional nucleotide-based substitution models. Correcting the apparent underestimation of branch lengths suggests substantially older origins for measles, Ebola, and avian influenza viruses. This work helps to reconcile some of the inconsistencies between molecular dating and other types of evidence concerning the age of viral lineages.

Key words: measles virus, rinderpest virus, Ebola virus, avian influenza virus, molecular clock, substitution rate, codon model, purifying selection.

Introduction

One of the most powerful forces shaping the human genome is adaptation to pathogens, and the pathogens that appear to have exerted the strongest influence are viruses (Worobey et al. 2007; Emerman and Malik 2010). Genes which bear the mark of some of the most potent selective forces detected in the genomes of humans and our relatives are directly related to combating RNA viruses (Meyerson and Sawyer 2011). Moreover, the ancient history of this association with RNA viruses is evidenced by a diverse array of defective viral remnants incorporated into vertebrate genomes, including mounting support for endogenization of RNA viruses other than Retroviridae (Gifford et al. 2008; Belyi et al. 2010; Gilbert and Feschotte 2010; Horie et al. 2010; Taylor et al. 2010). However, according to molecular dating analyses, many RNA viruses have extraordinarily recent origins (Holmes 2003a).

At first glance, a recent introduction of many RNA viruses into the human population is unsurprising. Epidemiological, historical, and phylogenetic approaches agree that some of the most notable RNA viruses (e.g., HIV-1, Worobey et al. 2008; influenza A virus, Taubenberger et al. 2005; Ebola virus [EBOV] Zaire, Walsh et al. 2005; and SARS coronavirus, Hon et al. 2008) emerged as zoonoses within the last century. As one looks further back in evolutionary time, however, specific inconsistencies arise. For example, defective remnants of ancient integrations of Filoviridae, the viral family containing the EBOV, are found in mammalian lineages that diverged tens of millions of years ago (Belyi et al. 2010; Taylor et al. 2010), even though molecular dating suggests a time of most recent common ancestor (tMRCA) for Filoviridae on the order of only thousands of years ago (Suzuki and Gojobori 1997). Another inconsistency can be seen in the zoonotic origin of measles virus (MeV) from rinderpest virus (RPV) and peste-des-petits ruminants virus (PPRV), two viruses capable of infecting large and small ruminants, respectively. MeV required at least two conditions to emerge: 1) Humans had to live in close proximity to RPV, which became commonplace only after the domestication of cattle over the last 10,000 years (Perkins 1969; Loftus et al. 1994; Beja-Pereira et al. 2006) and 2) humans needed societies with a population size above 250,000–500,000 to sustain the epidemic, which did not exist until about 5,000 years ago (Black 1966). Furthermore, the first unambiguous historical account of measles dates back to the ninth century (Rāzī 1848). Therefore, historical and epidemiological considerations, which place the tMRCA of MeV and RPV at thousands of years ago, are at odds with molecular dating analysis, which infers the tMRCA to be only hundreds of years ago (Furuse et al. 2010).

The same phylogenetic dating methods that provide convincing and reasonable tMRCA estimates for recent zoonotic transfers seem to be dramatically underestimating the age of older viral divergence events. Therefore, over longer periods of time, the extent of evolutionary change has been lost. In many viral genes, there is remarkable sequence conservation, indicating that purifying selection is a dominant evolutionary force, acting to maintain evidence of homology by preserving amino acid residues, while allowing nucleotide sequences to continue evolving

(Holmes 2003b; Edwards et al. 2006; Pybus et al. 2007). Although the effect of purifying selection on mutations in RNA viruses is becoming better understood (Belshaw et al. 2008), the importance of spatial and temporal variation in selection pressures has not been explicitly explored in the context of tMRCA estimation.

It has been well established that failing to account for site-to-site rate variation can lead to an underestimation of branch lengths due to repeated substitutions at rapidly evolving sites (Brown et al. 1982; Sullivan and Joyce 2005), and the use of models, which permit such variation (e.g., Yang 1993), have become standard. More recently, Suchard and Rambaut (2009) observed that synonymous substitutions appeared to saturate faster than could be handled by nucleotide models in mitochondrial DNA, which is subject to strong overall purifying selection. Over longer periods of evolutionary time (i.e., along long internal branches on a phylogenetic tree), evidence of evolution at the nucleotide level could be lost, which could bias tMRCA estimates towards younger dates (i.e., underestimate the length of these branches). If a substitution model can account for this evolution, we hypothesize that a comparative analysis of recent viral isolates could be used to make inference about ancient viral divergence events.

Our work addresses two questions. First, can purifying selection mask the ancient history of RNA viruses? And second, by using evolutionary models that account for selection, can we infer more realistic estimates of the ages of these viruses? We investigate these questions using two groups of viruses thought to be older than current molecular dating evidence would suggest: MeV/RPV/PPRV and EBOV. We also consider avian influenza virus (AIV)—another well-characterized viral lineage with a young inferred tMRCA and long internal branches between different serotypes (Chen and Holmes 2006, 2010). The results presented here demonstrate that not accounting for purifying selection may bias tMRCA estimation in RNA viruses towards more recent dates, and a degree of correction can be realized by employing more realistic codon-based substitution models, capable of partially accounting for the biasing effect of purifying selection.

Materials and Methods

Viral Gene Sequence Alignments

For MeV/RPV/PPRV, all available full-length nucleoprotein sequences with known years of isolation were downloaded from GenBank. Vaccine-associated sequences and MeV isolates implicated in subacute sclerosing panencephalitis, which experience hypermutation (Woelk et al. 2001, 2002), were excluded. The curated MeV/RPV/PPRV virus data set contained 145 sequences sampled between 1981 and 2007. Alignment was trivial due to a lack of insertions or deletions and was performed by eye. Next, EBOV full-length glycoprotein sequences with known years of isolation were downloaded from GenBank. Regions containing multiple reading frames (Volchkov et al. 1995; Sanchez et al. 1996) and the mucin-like region—which evolves extremely rapidly due to relaxed selective constraints (Wertheim and Worobey 2009b)—were removed. The curated EBOV data set contained 34 sequences sampled between 1976 and 2005. As before, alignment was trivial and performed by eye. The final alignment, AIV neuraminidase, was provided by Chen and Holmes (2010). It contained 270 sequences sampled between 1972 and 2005. All three alignments in NEXUS format can be downloaded from http://www.hyphy.org/wiki/codondating. Redundant sequences were excluded from the alignments for branch length comparisons and selection analyses.

Phylogenetic and Substitution Rate Inference

We used a Bayesian Markov chain Monte Carlo (BMCMC) method implemented in BEAST v1.5.4 for phylogenetic inference and tMRCA estimation (Drummond and Rambaut 2007). For each data set and substitution model, four independent BMCMC runs of 25 or 50 million generations were performed. A codon-based substitution model was implemented in BEAGLE (Suchard and Rambaut 2009). The first 10% of the generations were discarded as burn-in. All BMCMC analyses were performed using an uncorrelated lognormal relaxed molecular clock and a Bayesian skyline plot coalescent prior, which places the fewest demographic constraints on the analysis (Drummond et al. 2005, 2006). Tracer v1.5 was used to check for convergence and adequate mixing (i.e., estimated sample size >200 for all relevant parameters). Finally, the maximum clade credibility (MCC) phylogeny was identified and annotated using the posterior distribution of trees. Substitution rates and tMRCAs are reported as mean and 95% highest posterior density values.

Codon Substitution Models

An earlier study (Kosakovsky et al. 2005) developed a hierarchy of five models—all extensions of the Muse–Gaut probabilistic model of sequence evolution (Muse and Gaut 1994)—of differing complexities that incorporate site-to-site and lineage-to-lineage variation of synonymous (α) and nonsynonymous (β) substitution rates. We investigated the ability of these models to accurately infer branch lengths on our viral data sets. The five models are summarized below, with full details available in the original manuscript. Note that in all models, $E[\alpha] = 1$ to ensure identifiability.

Constant Rates: The baseline model, which extends the original Muse–Gaut evolutionary model by incorporating general nucleotide substitution biases (MG94 $\times$ REV). α and β rates are constant across sites and lineages.

Proportional: A direct analog to nucleotide + Γ_4 models; $\beta = \omega\alpha$ and α varies from site to site according to a three-bin general discrete distribution (GDD).

Nonsynonymous: A standard "selection" model, where α rates are constant and β rates are drawn from a 3-bin GDD.

Dual: A model, where both α and β rates vary from site to site, based on independent three-bin GDD distributions.

Lineage+Dual: In addition to site-to-site rate variation in α and β, some (or all) lineages are endowed with their own mean $E[\beta]/E[\alpha]$ ratios, to correct for "lineage-specific" effects of selection.

Unlike the original implementation, our version of the Lineage+Dual model treated certain interior lineages differently to better reflect the biological reality of the sample. Branches which separated clades containing different viral species (MeV/RPV/PPRV), subtypes (EBOV), and serotypes (AIV) were assigned separate $E[\beta]/E[\alpha]$ ratio parameters because they represent different timescales and selective regimes compared with terminal branches. Longer internal branches are expected to bear the mark of stronger purifying selection (Kosakovsky et al. 2006; Pybus et al. 2007). We considered two variations of this model: Either 1) all long internal lineages share the same ratio parameter, and the remaining branches share another global ratio parameter (two-rate model) or 2) each deep lineage possesses its own ratio parameter, and the remaining branches share a global ratio parameter (multirate model). For each data set, we selected the model with the best Akaike information criterion (AIC) score for further analysis. Note that we use the ratio of the means $E[\beta]/E[\alpha] = E[\beta]$ (due to the $E[\alpha] = 1$ identifiability constraint) instead of the mean of the ratio $E[\beta/\alpha] \equiv \omega$ because $\alpha = 0$ is possible under Dual and Lineage+Dual models, rendering the mean of the ratio infinite.

For each data set, we obtained $M = 1,000$ samples from the approximate joint distribution of model parameter estimates using a modified Latin hypercube sampling importance resampling scheme, described in detail elsewhere (Kosakovsky et al. 2010). The approach is meant to quickly obtain an approximate joint distribution of maximum likelihood parameter estimators. First, an area of parameter space to be sampled is defined by constructing a d-dimensional rectangle, in which each dimension represents a single model parameter, the corresponding coordinate interval is centered on the maximum likelihood estimate of the parameter, and the lower and upper bounds are determined by profile likelihood. Second, each coordinate interval is partitioned into $N = 1,000d$ subintervals of approximately equal probability, based on the asymptotic normal approximation to the likelihood surface. Third, N samples are drawn from the d-dimensional rectangle, using the Latin Hypercube scheme (i.e., each interval in every coordinate is sampled exactly once). Fourth, $M \ll N$ points are resampled based on their importance (normalized likelihood), following the procedure described elsewhere (Skare et al. 2003).

Variance in estimated branch lengths was computed using this sample. All codon-based analyses were performed in HyPhy v2.0020110301 (Kosakovsky et al. 2005).

Codon-Based Simulations

We simulated codon sequences along a single branch using the MG94 × REV codon substitution model (Kosakovsky et al. 2005) with site-specific β and α values inferred using a fixed effects likelihood method on internal branches (IFEL) (Kosakovsky et al. 2006) from each of the three viral data sets. In this method, three rates are inferred from each site: α_s, β_s^I, and β_s^T. Subscript s indicates the explicit dependence of substitution rates on the site, from which they are being estimated. α_s is the tree-wide synonymous rate, β_s^I is the nonsynonymous rate shared by all internal branches, and β_s^T is the nonsynonymous rate shared by all terminal branches. Only internal branches were used to generate empirical selection profiles because substitutions along tips in viral phylogenies are frequently deleterious and transient (Kosakovsky et al. 2006; Pybus et al. 2007). Furthermore, we were primarily interested in the effects of selection on long internal branches, and the IFEL profile (α_s and β_s^I inferred for each site) was meant to recapitulate the evolutionary process along an "average" internal branch. Simulations were initialized using ancestral nucleotide sequences inferred with marginal maximum likelihood reconstruction (Yang et al. 1995) in HyPhy, using the MG94 × REV codon substitution model on the MCC phylogeny obtained from BMCMC general time reversible (GTR) + Γ_4 analyses. The marginal ancestral sequence was used to provide a realistic starting point for our simulations. Ten thousand replicates were generated for each branch length ranging from 0.01 to 100 expected substitutions per nucleotide site and analyzed under the GTR + Γ_4 and Dual substitution models. We chose to simulate branches based on expected substitutions per nucleotide site, instead of per unit time, because of the difficulty in standardizing time for variable rate parameters (i.e., α_s, β_s^I, and substitution rate) in a reversible model; the expected number of substitutions divided by the substitution rate can be a proxy for time.

Results

Estimating the Age of Viral Lineages

Using BMCMC analysis, we inferred the substitution rate and root age under a standard nucleotide substitution model (GTR + Γ_4) for each of our three viral data sets: MeV/RPV/PPRV nucleoprotein, EBOV glycoprotein, and AIV neuraminidase (table 1). The viral lineages examined here had tMRCAs ranging from hundreds to several thousand years before present. All three viruses exhibited rapid substitution rates, on par with previous estimates for related RNA viruses (Duffy et al. 2008).

Our inferred substitution rate for the MeV/RPV/PPRV nucleoprotein of 9.65×10^{-4} (6.00×10^{-4}–1.47×10^{-3}) substitutions/site/year under a GTR + Γ_4 model was over 50% faster than the rate recently reported by Furuse et al. (2010): 6.02×10^{-4} (3.62×10^{-4}–8.76×10^{-4}) substitutions/site/year. Therefore, we inferred a substantially younger tMRCA for MeV and RPV: 1490 CE (1130–1810 CE), instead of 1171 CE (678–1612 CE). We consider our rate estimates more reliable for three reasons. First, our values are in close agreement with previous substitution rate estimates in MeV: 8.69×10^{-4} (5.89×10^{-4}–1.13×10^{-3}) substitutions/site/year (Pomeroy et al. 2008). Second, Furuse et al. (2010) included both MeV and RPV vaccine strains in their BMCMC dating analysis (Rota et al. 1994; Baron et al. 1996),

Table 1. Mean tMRCA and 95% highest posterior density for the root of viral lineages inferred under different evolutionary models.

Evolutionary model	MeV/RPV/PPRV	EBOV	AIV
GTR	333 (165–528)	819 (274–1,514)	333 (268–396)
GTR + Γ_4	667 (333–1,062)	1,492 (421–2,869)	1,265 (965–1,618)
GTR + Γ_4 (third position excluded)	285 (110–515)	498 (67–1,118)	1,493 (938–2,089)
SRD06	646 (288–1,051)	1,358 (378–2,592)	1,103 (855–1,378)
Whelan and Goldman + Γ_4	265 (94–477)	1,243 (87–2,948)	942 (619–1,334)
GY94 + Γ_4	698 (353–1,088)	2,247 (751–4,170)	1,243 (972–1,528)

despite evidence that the inclusion of vaccine strains can lead to bias in rate estimation (Bush et al. 2000; Wertheim 2010). Third, Furuse et al. (2010) fitted an exponential growth coalescent prior to viruses whose effective population size has been declining—due to eradication efforts against both MeV and RPV (Moss 2009; Normile 2010). Remarkably, our results are in even greater conflict with the historical documentation of measles (Rāzī 1848), further confirming our conjecture that traditional nucleotide substitution models can underestimate the age of ancient viral divergence events.

We then explored whether alternative evolutionary models that, to varying extents, account for codon structure were able to recover the expected older tMRCAs for these viruses (table 1). The simplest approach, excluding the third codon position in a GTR + Γ_4 model, has been used to remove synonymous sites that might have experienced saturation (Worobey et al. 2010). A more sophisticated approach, the SRD06 model (Shapiro et al. 2006), allows first and second codon positions to have a different transition/transversion ratio and Γ_4 shape parameter from the third position. Amino acid substitution models have also been successfully implemented to estimate the age of RNA viruses (Zlateva et al. 2005; Wertheim et al. 2009). Finally, we evaluated an available codon-based substitution model, GY94 (Goldman and Yang 1994). Generally, these alternate models either produced younger tMRCAs (e.g., GTR + Γ_4 excluding third positions and Whelan and Goldman + Γ_4) or tMRCAs that were indistinguishable from GTR + Γ_4 (e.g., SRD06 and GY94 + Γ_4). A possible exception was analysis of the EBOV data set with GY94 + Γ_4, which produced a root tMRCA that was 50% older than the GTR + Γ_4 estimate. Finally, as a point of comparison, we also investigated how removing site-to-site rate variation affected dating inference by using a GTR model (Tavaré 1986); ignoring rate variation invariably produced younger tMRCAs.

Purifying Selection Leads to Underestimation of Branch Lengths

We simulated codon sequences along a single branch under a substitution model with site-to-site nonsynonymous rate variation and inferred the length of the branch using GTR + Γ_4. An increase in the proportion of sites simulated under purifying selection (i.e., $\beta_s = 0$) resulted in shorter branches inferred by a GTR + Γ_4 model (fig. 1*A*). Sufficiently strong purifying selection could theoretically lead to underestimates of branch lengths by an order of magnitude because purifying selection slows down the rate of evolution relative to the synonymous rate. At the extreme ($\beta_s = 0$, branch length $\rightarrow \infty$), one obtains perfectly conserved amino acid sequences and completely saturated synonymous substitutions. For sequences with 10% of sites under strict conservation, the nucleotide estimator approached the asymptote of 7.1 substitutions/site, even when the simulated branch length was increased to 100; the branch length estimate was accurate for up to 5 substitutions/site. For sequences with no nonsynonymous substitutions (i.e., 100% conservation), saturation occurred much sooner, around 0.25 substitutions/site, and the corresponding asymptote was 0.36 substitutions/site.

The nucleotide substitution model (GTR + Γ_4) underestimated the length of branches simulated under empirical (IFEL) selection regimes inferred from the three viral data sets (fig. 1*B*–1*D*); branch lengths for sequences simulated under neutral selection (i.e., $\alpha_s = \beta_s = 1$) were reliably inferred by GTR + Γ_4. The proportion of sites under strong negative selection (i.e., $\beta_s < \alpha_s$ with an IFEL P value $\leqslant 0.05$) differed markedly among the three data sets. MeV/RPV/PPRV nucleoprotein and EBOV glycoprotein genes experienced moderate levels of purifying selection along internal branches: 50% and 45% of sites under strong purifying selection, respectively. The AIV neuraminidase gene evolved under much stronger selective constraints, with 89% of sites under strong purifying selection. We note that the power to detect selection using IFEL increases with the size of the alignment (Kosakovsky et al. 2006); hence, we would expect more sites evolving under purifying selection to be correctly detected as such in the larger AIV data set. The simulated codon saturation curves varied from virus to virus, indicating that the particular selective regime, amino acid composition, and nucleotide substitution model of each viral gene have a substantial influence on the ability to accurately infer branch lengths. Although underestimation is more pronounced along longer branches (e.g., about half the true length for 1 substitution/site), not accounting for selection can lead to underestimation of branch lengths even for relatively short branches (e.g., 0.2–0.5 substitutions/site).

We then inferred the length of branches simulated under the empirical selective regimes using the Dual codon model, which accounts for variation in site-to-site selective pressures (fig. 2). Although the Dual model slightly underestimated lengths for the longest simulated branches, the bias was an order of magnitude less than under GTR + Γ_4. Furthermore, similar behavior was observed for the longest branches simulated under neutrality ($\alpha_s = \beta_s$) and

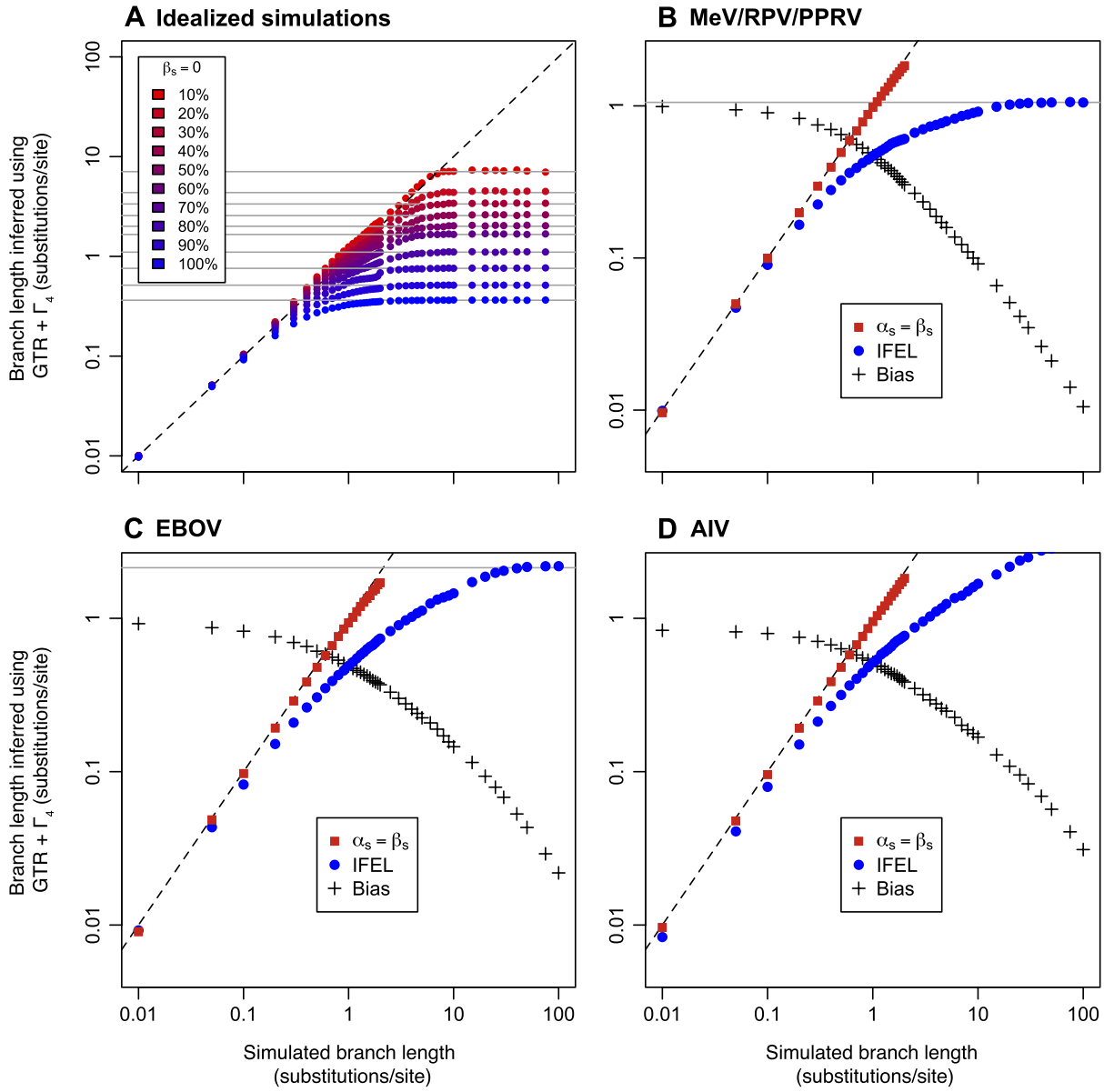

FIG. 1. Lengths of single branches simulated under a codon substitution model with variable selection pressures are underestimated when inferred under a GTR + Γ_4 nucleotide substitution model. All quantities are means from 10,000 replicates. (*A*) Ancestral sequence (1,000 codons) was random (i.e., all codons equiprobable) with varied proportion of sites under strong purifying selection ($\beta_s = 0$) and sites evolving neutrally ($\alpha_s = \beta_s$). (*B–D*) Branch lengths were inferred using GTR + Γ_4 on single branches simulated under neutral selection regimes ($\alpha_s = \beta_s$) and empirical site-by-site substitution rate (IFEL) profiles for MeV/RPV/PPRV, EBOV, and AIV. Black crosses represent the degree of bias, defined as the ratio between the true branch length and the one inferred under GTR + Γ_4 on IFEL sequences. Horizontal gray lines show saturation asymptotes and diagonal dashed line depicts the behavior of an unbiased estimator.

inferred using GTR + Γ_4 (fig. 1*B*–1*D*), suggesting that both models perform equivalently, as expected. Clearly, not accounting for purifying selection can lead to dramatic underestimation of branch lengths; this effect is exacerbated for longer branches.

Effect of Evolutionary Models on Branch Lengths

It is well known that standard nucleotide models differ in their sensitivity to multiple substitutions at the same site (Sullivan and Joyce 2005). We explored how various methods of modeling rate variation affected branch length inference in our three viral data sets, using a fixed MCC tree from the GTR + Γ_4 BMCMC analyses. First, we examined two extremes among nucleotide substitution models that do not adjust for rate variation across sites: JC69 (Jukes and Cantor 1969) and GTR. JC69 assumes a single substitution rate and equal base frequencies, whereas GTR allows each class of nucleotide substitution to occur at a unique rate and estimates base frequencies from the data. As expected, failure to account for site-to-site rate variation led to severe underestimation of longer branches in all three viral data sets (JC69 or GTR vs. GTR + Γ_4, fig. 3), reaffirming

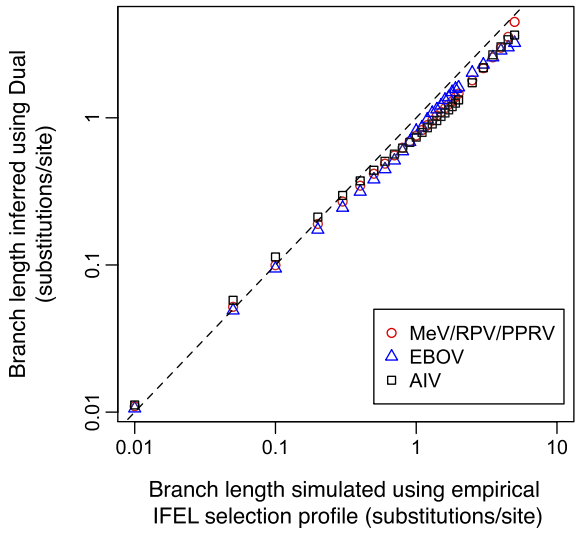

FIG. 2. Branch lengths inferred under a Dual model provide reliable estimates of branches simulated under variable selection pressures (i.e., empirical IFEL profiles) for MeV/RPV/PPRV, EBOV, and AIV. The diagonal dashed lines depict the behavior of an unbiased estimator.

the importance of modeling site-to-site rate variation to accurate inference. Allowing for multiple substitution rates and empirical base frequencies had a negligible impact on branch length estimation (fig. 3). Underestimation of long branches likely explains why the GTR model inferred younger root tMRCAs for all all three viral data sets (table 1).

We hypothesized that codon substitution models, which explicitly account for the differences between synonymous and nonsynonymous substitutions, would permit a more accurate estimation of sequence divergence well past the point a standard nucleotide model would reach saturation. Therefore, it may be possible to use RNA virus genes, which evolve extremely rapidly (e.g., 0.0001–0.001 substitutions/site/year), to estimate ancient viral divergence events using codon substitution models. A crude approximation of a codon model, SRD06, produced branch lengths that were essentially the same as those from the GTR + Γ_4 model (fig. 3), which may explain why BMCMC tMRCA inference using these two models was so similar. The inference under the GY94 + Γ_4 codon model resulted in longer internal branch lengths only for EBOV (fig. 3), which was in agreement with our BMCMC tMRCA inference using this model.

Based on the simulations along a single branch, we anticipated that longer branches would be disproportionately affected by site-to-site variation in selection pressures. Therefore, we investigated how extensions of the MG94 model that incorporate β, α, and lineage-specific rate variation affected branch length inference (table 2). For MeV and AIV, the Lineage+Dual (two-rate) model that allowed longer internal branches to share their own ratio $E[\beta]/E[\alpha]$ provided the best fit to the data. For EBOV, a more complicated (eight-rate) model, in which each intersubtype branch and the long terminal branches leading to Côte d'Ivoire and Bundibugyo had their own rates, was fit because $E[\beta]/E[\alpha]$ varied dramatically among the intersubtype branches. The inclusion of variation in β and α in the Nonsynonymous and Dual models resulted in notable increases in the cumulative length of deep internal branches (table 2; supplementary fig. S1, Supplementary Material online). In all three viral data sets, however, the most substantial differences were seen with the Lineage+Dual model, which produced lengths for long internal branches that were substantially greater than those inferred under GTR + Γ_4 (fig. 3). Importantly, the lengths of the more recent intraspecies/subtype/serotype branches were relatively unchanged between these two models for all three data sets; this observation confirms that when the phylogeny is comprised of relatively short branches, it is appropriate to rely on common approximations, such as the GTR + Γ_4 model. This increase in the length of only deep internal branches was likely due to the dramatically stronger purifying selection which we inferred along these lineages (table 2). In AIV, for instance, $E[\beta]/E[\alpha]$ was two orders of magnitude lower on deep interior branches, compared with the rest of the tree.

One outcome of including variation in selection pressures across branches into the evolutionary model was that several branches were inferred to have essentially infinite lengths in the EBOV and AIV trees. The inference of an infinite branch length is likely caused by the complete saturation of synonymous substitutions along the branch in question; even after accounting for variation in α_s and β_s, the true branch length was inestimable. In the EBOV phylogeny, complete saturation was observed on the branch leading to EBOV Sudan and on the branch connecting EBOV Reston/Sudan to EBOV Zaire/Côte d'Ivoire/Bundibugyo (supplementary fig. S2, Supplementary Material online). In the AIV phylogeny, each of the nine neuraminidase serotypes naturally found in avian hosts was separated by branches that experienced saturation at synonymous sites under the Lineage+Dual model (supplementary fig. S3, Supplementary Material online).

The Latin hypercube resampling scheme suggested relatively narrow variance in the expansion under the Lineage+Dual model. The total increase in MeV/RPV/PPRV tree relative to GTR + Γ_4 was 1.93 (approximate 95% confidence interval: 1.76–2.16). Due to the inference of branch lengths that experienced saturation in EBOV and AIV, the overall increase in tree length was more dramatic under the Lineage+Dual model, relative to GTR + Γ_4: 32.14 (approximate 95% confidence interval: 12.14–32.22) for EBOV and 107.34 (approximate 95% confidence interval: 101.06–117.36) for AIV.

Implications for the Age of Viral Lineages

It is clear that purifying selection can obscure the ancient evolution of the RNA viruses examined here. A comparison of the MeV/RPV/PPRV phylogeny optimized under GTR + Γ_4 and Lineage+Dual (two-rate) models showed that the elongation of branches occurred along the deep internal branches, leaving the relationships among recently divergent lineages within viral species relatively unchanged (fig. 4). The depth of the split between MeV and RPV under a GTR + Γ_4 model was 0.4921 substitutions/site, which, according to the BMCMC analysis, occurred in the year 1483

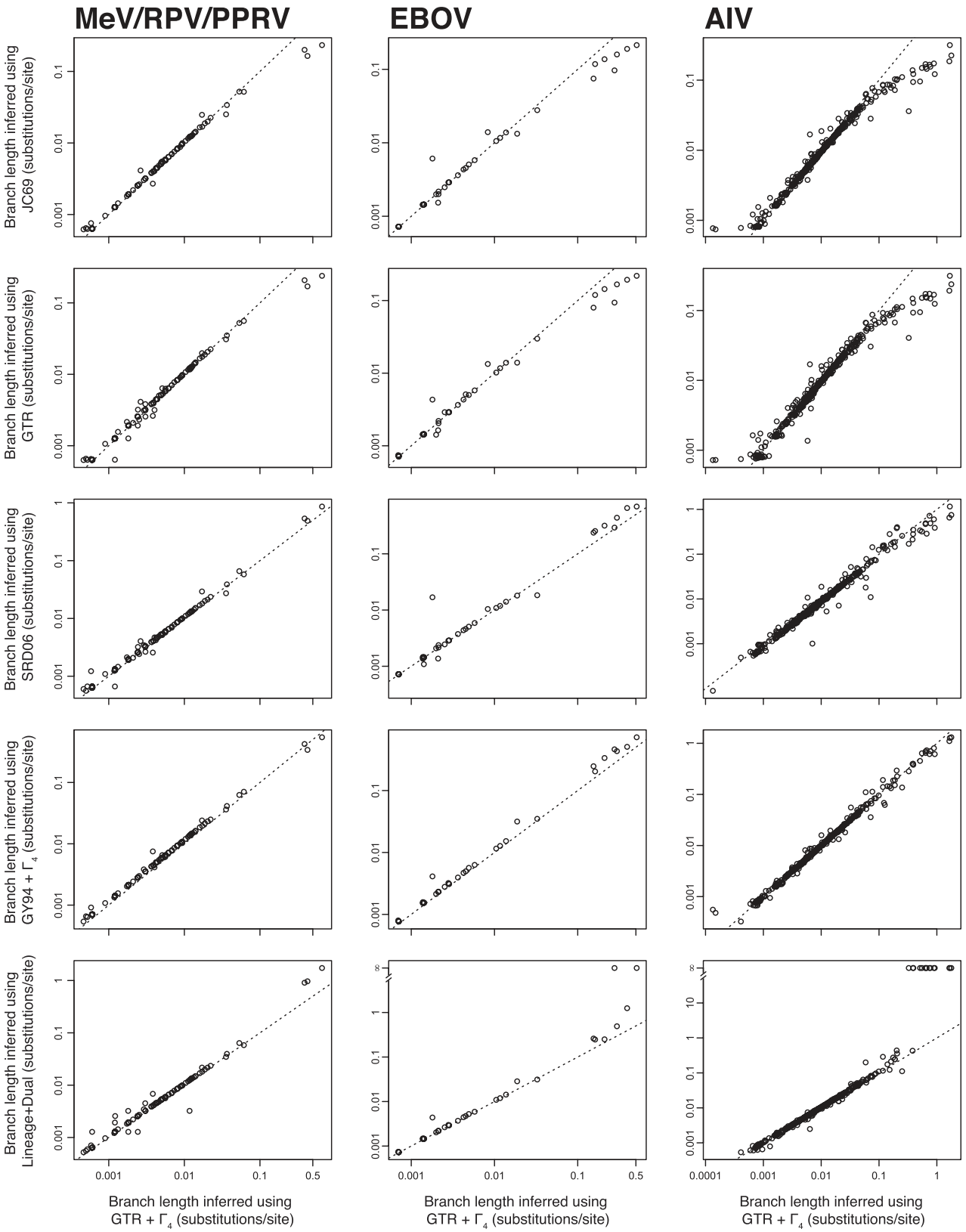

FIG. 3. Long branches are disproportionately affected by evolutionary models that differ in their treatment of rate variation. Each datapoint represents the length of a single branch of the MeV/RPV/PPRV, EBOV, or AIV phylogeny inferred under GTR + Γ_4 and an alternate evolutionary model. The extreme of the y axis represents infinite branch lengths under the Lineage+Dual model for EBOV and AIV. Dashed lines are an $x = y$ reference.

Table 2. Goodness of fit for various codon models and the effect of model choice on the estimates of the branch lengths of deep and recent lineages.

Taxa	Model[a]	Log *L*	AIC[b]	*T*[c]		$E[\beta]/E[\alpha]$[d]	
				Deep	Recent	Deep	Recent
MeV/RPV/PPRV	Constant	−12207.97	24909.95	1.29	0.91	0.10	
	Proportional	−12122.06	24746.11	1.46	0.93	0.10	
	Nonsynonymous	−11928.55	24359.11	1.93	0.92	0.12	
	Dual	−11919.35	24348.70	2.00	0.93	0.12	
	Lineage+Dual (two rate)	−11900.7	24313.40	3.61	0.92	0.04	0.14
	Lineage+Dual (four rate)	−11900.1	24316.20	3.95	0.92	000.03–0.06	0.14
EBOV	Constant	−6762.76	13757.52	2.37	0.17	0.05	
	Proportional	−6720.47	13676.94	3.00	0.17	0.04	
	Nonsynonymous	−6682.31	13600.62	2.56	0.17	0.06	
	Dual	−6679.79	13599.57	2.68	0.17	0.06	
	Lineage+Dual (two rate)	−6638.5	13420.99	4.84	0.17	0.03	0.18
	Lineage+Dual (eight rate)	−6631.24	13418.49	69.4	0.17	0.0005–0.05	0.18
AIV	Constant	−44860.13	90794.26	11.96	8.8	0.05	
	Proportional	−44483.52	90049.05	12.80	9.04	0.05	
	Nonsynonymous	−43862.45	88806.90	20.92	9.08	0.05	
	Dual	−43730.4	88550.9	20.98	9.2	0.05	
	Lineage+Dual (two rate)	−43711.00	88514.0	2230.9	9.2	0.0004	0.05
	Lineage+Dual (16 rate)	−43710.97	88541.9	2278.23	9.2	0.0003–0.0004	0.05

NOTE.—ALC, Akaike information criterion.
[a]The number of lineages (selected a priori) with their own $E[\beta]/E[\alpha]$ for the Lineage+Dual models are shown in parentheses.
[b]The best fitting model for each data set is highlighted in boldface.
[c]*T* shows the cumulative length of branches classified as recent or deep lineages a priori, measured in the expected number of substitutions per nucleotide site.
[d]$E[\beta]/E[\alpha]$ reports the ratio of means of the expected nonsynonymous to expected synonymous rates (similar to ω for each model). For Lineage+Dual models, the values are stratified by branch class and ranges are reported when appropriate.

CE (1162–1777 CE). The depth of this split using a Lineage+Dual (two-rate) model was 1.0036 substitutions/site: 2.04 times deeper. A simple extension of the substitution rate inferred from a MeV-only data set of 9.06×10^{-4} (7.12×10^{-4}–1.17×10^{-3}) substitutions/site/year (likely preferable to the inferred MeV/RPV/PPRV substitution rate which relied on demonstrably biased estimates of internal branch lengths) would place the tMRCA of MeV and RPV in the year 899 CE (597–1144 CE). This extrapolation is meant as a rather crude approximation of the age of this divergence event; nevertheless, these dates are more likely to be closer to the true split between MeV and RPV than previous estimates and are not inconsistent with recorded history of measles.

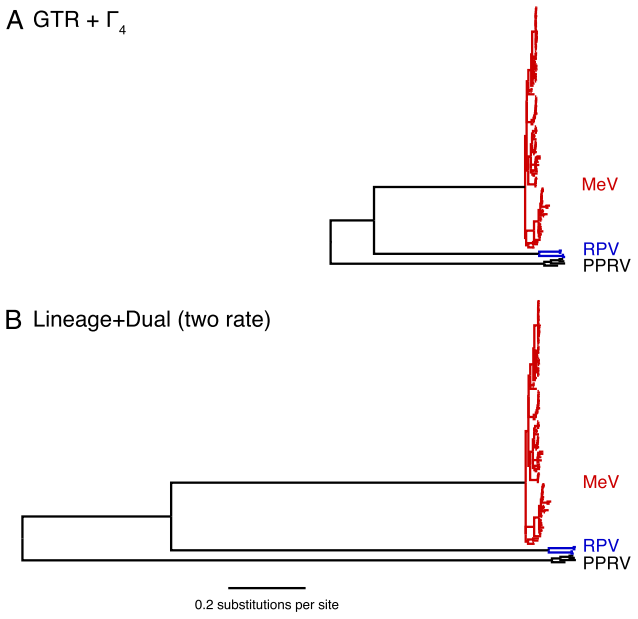

FIG. 4. MCC phylogeny for MeV, RPV, and PPRV. Branch lengths were optimized under (*A*) GTR + Γ_4 and (*B*) Lineage+Dual models. The Lineage+Dual model branch lengths were estimated assuming two different sets of synonymous and nonsynonymous substitution rates: one for short branches and another for long internal branches. Both trees are shown on the same scale.

The degree of synonymous saturation observed along the EBOV and AIV phylogenies indicate an inability to reliably infer tMRCAs for these viruses. Too many sites have sunk beyond the evolutionary horizon. Nevertheless, these estimates could provide meaningful minimum bounds for the tMRCAs. Thus, for both EBOV and AIV, we applied the mean substitution rate inferred in the BMCMC GTR + Γ_4 analyses to the Lineage+Dual phylogenies. This approach suggested that the minimum tMRCA estimates for EBOV and AIV are approximately 70,000 and 200,000 years ago, respectively. The actual tMRCAs may be much older.

Discussion

Our results suggest that the ancient age of RNA viruses may be partially masked behind a veil of purifying selection. The same forces of purifying selection that maintain evidence of protein sequence homology over great evolutionary distances also truncate long ancestral branches deep within phylogenetic trees. We observed this pattern

in three different groups of rapidly evolving negative-sense RNA viruses: MeV/RPV/PPRV, EBOV, and AIV. Estimating branch lengths under a codon-based substitution model that accounts for spatial and temporal variation in selection pressures yielded phylogenetic trees that were more than twice as long as those obtained under standard nucleotide models. Modeling synonymous, nonsynonymous, and lineage-specific rate variation indicates that current estimates of the age of these, and possibly other, viral lineages may be dramatically underestimated. Codon models used in this study do not circumvent the issue of substitutional saturation but merely extend the horizon further back in evolutionary time; the application of more complex and biologically realistic models is likely to extend these branches even further into the past. Eventually, however, even the most realistic models are likely to fail and external information, such as biogeography and homology to endogenous viral elements, must be brought into consideration to estimate truly ancient events (Katzourakis et al. 2009; Katzourakis and Gifford 2010).

The precise age of measles in the human population remains elusive. The domestication of cattle beginning 10,000 years ago provided the necessary exposure to RPV, and the development of agriculture allowed human populations to reach sizes necessary to sustain an epidemic. However, the historical record of measles is ambiguous until the ninth century, when Rhazes (a Persian physician) outlined the criteria for differentiating between measles and smallpox (Rāzī 1848). Rhazes discusses both ailments as immemorial, indicating that both diseases predated him by many generations. Although there are historical records of earlier plagues that could be interpreted as measles, notably in eighth century France during the battle of Tours (Rolleston 1937), it is also possible that this plague could have been smallpox or another disease with similar presentation. Even after Rhazes description, Western medicine still confounded measles, smallpox, and scarlet fever until the 17th century (Rolleston 1937). In light of this confusion, the lack of a definitive description of measles before Rhazes (e.g., Galen, an ancient Roman physician, described smallpox but not measles) does not establish that the virus was absent (McNeill 1976). It is plausible that MeV might have not entered the human population until the first millennium of the Common Era, as our analysis suggests. Alternatively, MeV could have emerged thousands of years ago, and the evolutionary models employed here are too crude to recover the true age of the virus. Regardless of which scenario is correct, accounting for variable selective pressures is an important step in revealing the ancient history of RNA viruses.

The different timescales affecting mutation rates and substitution rates make it difficult to extrapolate population-based estimates of rates over a long evolutionary time (Ho et al. 2005); however, it is unlikely that purifying selection alone can account for this difference between short- and long-term evolutionary rates (Woodhams 2006). Furthermore, short-term estimates of viral substitution rates (inferred from population-based rate estimates at the tips of phylogenies) are often found to be several orders of magnitude faster than long-term estimates of substitution rates (inferred from external calibrations located deeper in the phylogeny). This inconsistency has been reported for hepatitis B virus (Zhou and Holmes 2007; Gilbert and Feschotte 2010) and simian immunodeficiency virus (Wertheim and Worobey 2009a; Worobey et al. 2010), suggesting that long-term substitution rates can be orders of magnitude lower than short-term substitution rates. An alternative, and more parsimonious, explanation is that a single substitution rate (albeit one possibly slower than that inferred using short-term population-based data) predominates throughout the history of these viruses. And our inability to accurately estimate branch lengths creates the appearance of dramatically lower substitution rates deep in the phylogenetic tree. Although the methods presented here do not correct for branch lengths on the order needed to reconcile short-term substitution rates with deep calibrations, they provide a glimpse at the missing evolution that most current methods of phylogenetic inference fail to capture.

Although we are unable to provide a full remedy to the problem of underestimated branch lengths due to purifying selection, our results do point to a couple of guidelines when inferring tMRCAs in RNA viruses. First, the inference of substantially different selective regimes on longer internal branches compared with shorter shallower branches (e.g., using Lineage+Dual, Kosakovsky et al. 2005, or Free Ratio, Yang 1998, codon models) appears to be a sign that older tMRCAs may be underestimated. Second, if the synonymous substitution rate approaches saturation along a branch or group of branches, it is likely that the tMRCA cannot be reliably inferred. If either of these patterns is encountered, the inferred tMRCA estimates should be interpreted with caution. Rapid advances in cheap and accessible computational power will undoubtedly move the fields of paleovirology and molecular dating towards increasingly more realistic evolutionary models that account for variation in selective regimes (Suchard and Rambaut 2009). Our results provide compelling evidence that such a movement should be accelerated.

In this study, we demonstrate the need to include relevant biological and evolutionary forces in substitution models. It was not until we permitted different evolutionary regimes for different branches in the tree that we saw the most dramatic changes in branch length estimates. Current codon models take a very simplistic mechanistic view of the action of purifying selection, and we expect that incorporating processes such as directional selection (Seoighe et al. 2007), toggling selection (Delport et al. 2008), and residue-specific (Doron-Faigenboim and Pupko 2007) or site-specific substitution (Lartillot and Philippe 2004) biases can further refine tMRCA estimates.

It is still unclear how correcting for the effects of purifying selection may affect dating and branch length estimation in other types of viruses (e.g., DNA viruses) and cellular organisms. The bias we observed in the estimation of branch lengths appears to be related to the intensity of selection and the length of the branches in question.

The relative importance of these factors in understanding the evolutionary history of other taxa remains to be seen.

Supplementary Material

Supplementary figs. S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank the Associate Editor Alexei Drummond, Jeff Thorne, and two anonymous reviewers for their valuable comments. This research was supported in part by the National Institutes of Health (AI47745, AI57167, AI74621, and GM093939); the Joint Division of Mathematical Sciences/National Institute of General Medical Sciences Mathematical Biology Initiative through Grant NSF-0714991; the University of California University-wide AIDS Research Program (grant number IS02-SD-701); a University of California, San Diego Center for AIDS Research/National Institute of Allergy and Infectious Disease Developmental Award (AI36214 to S.L.K.P.); and a National Institutes of Health Training Fellowship (AI43638 to J.O.W.).

References

Baron MD, Kamata Y, Barras V, Goatley L, Barrett T. 1996. The genome sequence of the virulent Kabete 'O' strain of rinderpest virus: comparison with the derived vaccine. *J Gen Virol*. 77(Pt 12): 3041–3046.

Beja-Pereira A, Caramelli D, Lalueza-Fox C, et al. (27 co-authors). 2006. The origin of European cattle: evidence from modern and ancient DNA. *Proc Natl Acad Sci U S A*. 103:8113–8118.

Belshaw R, Gardner A, Rambaut A, Pybus OG. 2008. Pacing a small cage: mutation and RNA viruses. *Trends Ecol Evol*. 23:188–193.

Belyi VA, Levine AJ, Skalka AM. 2010. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. *PLoS Pathog*. 6:e1001030.

Black FL. 1966. Measles endemicity in insular populations: critical community size and its evolutionary implication. *J Theor Biol*. 11: 207–211.

Brown WM, Prager EM, Wang A, Wilson AC. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J Mol Evol*. 18:225–239.

Bush RM, Smith CB, Cox NJ, Fitch WM. 2000. Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *Proc Natl Acad Sci U S A*. 97: 6974–6980.

Chen R, Holmes EC. 2006. Avian influenza virus exhibits rapid evolutionary dynamics. *Mol Biol Evol*. 23:2336–2341.

Chen R, Holmes EC. 2010. Hitchhiking and the population genetic structure of avian influenza virus. *J Mol Evol*. 70:98–105.

Delport W, Scheffler K, Seoighe C. 2008. Frequent toggling between alternative amino acids is driven by selection in HIV-1. *PLoS Pathog*. 4:e1000242.

Doron-Faigenboim A, Pupko T. 2007. A combined empirical and mechanistic codon model. *Mol Biol Evol*. 24:388–397.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol*. 4:e88.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*. 7:214.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*. 22:1185–1192.

Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet*. 9: 267–276.

Edwards CTT, Holmes EC, Pybus OG, Wilson DJ, Viscidi RP, Abrams EJ, Phillips RE, Drummond AJ. 2006. Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. *Genetics* 174:1441–1453.

Emerman M, Malik HS. 2010. Paleovirology-modern consequences of ancient viruses. *Plos Biol*. 8.

Furuse Y, Suzuki A, Oshitani H. 2010. Origin of measles virus: divergence from rinderpest virus between the 11th and 12th centuries. *Virol J*. 7:52.

Gifford RJ, Katzourakis A, Tristem M, Pybus OG, Winters M, Shafer RW. 2008. A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. *Proc Natl Acad Sci U S A*. 105:20362–20367.

Gilbert C, Feschotte C. 2010. Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biol*. 8:e1000495.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*. 11: 725–736.

Ho SYW, Phillips MJ, Cooper A, Drummond AJ. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol*. 22:1561–1568.

Holmes EC. 2003a. Molecular clocks and the puzzle of RNA virus origins. *J Virol*. 77:3893–3897.

Holmes EC. 2003b. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. *J Virol*. 77:11296–11298.

Hon CC, Lam TY, Shi ZL, Drummond AJ, Yip CW, Zeng F, Lam PY, Leung FCC. 2008. Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. *J Virol*. 82:1819–1826.

Horie M, Honda T, Suzuki Y, et al. (11 co-authors). 2010. Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* 463:84–87.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. Vol. 3. New York: Academic Press. pp. 21–132.

Katzourakis A, Gifford RJ. 2010. Endogenous viral elements in animal genomes. *PLoS Genet*. 6:e1001191.

Katzourakis A, Gifford RJ, Tristem M, Gilbert MTP, Pybus OG. 2009. Macroevolution of complex retroviruses. *Science* 325:1512.

Kosakovsky Pond SL, Frost SDW, Grossman Z, Gravenor MB, Richman DD, Brown AJL. 2006. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Comput Biol*. 2:e62.

Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.

Kosakovsky Pond SL, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol*. 22:2375–2385.

Kosakovsky Pond SL, Scheffler K, Gravenor MB, Poon AFY, Frost SDW. 2010. Evolutionary fingerprinting of genes. *Mol Biol Evol*. 27: 520–536.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. 21:1095–1109.

Loftus RT, MacHugh DE, Bradley DG, Sharp PM, Cunningham P. 1994. Evidence for two independent domestications of cattle. *Proc Natl Acad Sci U S A*. 91:2757–2761.

McNeill WH. 1976. Plagues and peoples. 1st ed. Garden City (NY): Anchor Press.

Meyerson NR, Sawyer SL. 2011. Two-stepping through time: mammals and viruses. *Trends Microbiol*.
Moss WJ. 2009. Measles control and the prospect of eradication. *Curr Top Microbiol Immunol*. 330:173–189.
Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*. 11:715–724.
Normile D. 2010. Animal science. rinderpest, deadly for cattle, joins smallpox as a vanquished disease. *Science* 330:435.
Perkins D Jr. 1969. Fauna of Catal Hüyük: evidence for early cattle domestication in Anatolia. *Science* 164:177–179.
Pomeroy LW, Bjørnstad ON, Holmes EC. 2008. The evolutionary and epidemiological dynamics of the paramyxoviridae. *J Mol Evol*. 66:98–106.
Pybus OG, Rambaut A, Belshaw R, Freckleton RP, Drummond AJ, Holmes EC. 2007. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Mol Biol Evol*. 24:845–852.
Rāzī ABMiZ. 1848. A treatise on the small-pox and measles. London: Sydenham Society.
Rolleston J. 1937. The history of the acute exanthemata. London: William Heinemann (Medical Books) LTD.
Rota J, Wang Z, Rota P, Bellini W. 1994. Comparison of sequences of the H, F, and N coding genes of measles-virus vaccine strains. *Virus Res*. 31:317–330.
Sanchez A, Trappier SG, Mahy BW, Peters CJ, Nichol ST. 1996. The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing. *Proc Natl Acad Sci U S A*. 93:3602–3607.
Seoighe C, Ketwaroo F, Pillay V, et al. (11 co-authors). 2007. A model of directional selection applied to the evolution of drug resistance in HIV-1. *Mol Biol Evol*. 24:1025–1031.
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.
Skare Ø, Bølviken E, Holden L. 2003. Improved sampling-importance resampling and reduced bias importance sampling. *Scand J Statist*. 30:719–737.
Suchard MA, Rambaut A. 2009. Many-core algorithms for statistical phylogenetics. *Bioinformatics* 25:1370–1376.
Sullivan J, Joyce P. 2005. Model selection in phylogenetics. *Annu Rev Ecol Evol Syst*. 36:445–466.
Suzuki Y, Gojobori T. 1997. The origin and evolution of Ebola and Marburg viruses. *Mol Biol Evol*. 14:800–806.
Taubenberger JK, Reid AH, Lourens RM, Wang R, Jin G, Fanning TG. 2005. Characterization of the 1918 influenza virus polymerase genes. *Nature* 437:889–893.
Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci*. 17:57–86.
Taylor DJ, Leach RW, Bruenn J. 2010. Filoviruses are ancient and integrated into mammalian genomes. *BMC Evol Biol*. 10:193.
Volchkov VE, Becker S, Volchkova VA, Ternovoj VA, Kotov AN, Netesov SV, Klenk HD. 1995. GP mRNA of Ebola virus is edited by the Ebola virus polymerase and by T7 and vaccinia virus polymerases. *Virology* 214:421–430.
Walsh P, Biek R, Real L. 2005. Wave-like spread of Ebola Zaire. *PloS Biol*. 3:1946–1953.
Wertheim JO. 2010. The re-emergence of H1N1 influenza virus in 1977: a cautionary tale for estimating divergence times using biologically unrealistic sampling dates. *PLoS One* 5:e11184.
Wertheim JO, Tang KFJ, Navarro SA, Lightner DV. 2009. A quick fuse and the emergence of taura syndrome virus. *Virology* 390:324–329.
Wertheim JO, Worobey M. 2009a. Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Comput Biol*. 5:e1000377.
Wertheim JO, Worobey M. 2009b. Relaxed selection and the evolution of RNA virus mucin-like pathogenicity factors. *J Virol*. 83:4690–4694.
Woelk CH, Jin L, Holmes EC, Brown DW. 2001. Immune and artificial selection in the haemagglutinin (H) glycoprotein of measles virus. *J Gen Virol*. 82:2463–2474.
Woelk CH, Pybus OG, Jin L, Brown DWG, Holmes EC. 2002. Increased positive selection pressure in persistent (SSPE) versus acute measles virus infections. *J Gen Virol*. 83:1419–1430.
Woodhams M. 2006. Can deleterious mutations explain the time dependency of molecular rate estimates? *Mol Biol Evol*. 23:2271–2273.
Worobey M, Bjork A, Wertheim JO. 2007. Point, counterpoint: the evolution of pathogenic viruses and their human hosts. *Annu Rev Ecol Evol Syst*. 38:515–540.
Worobey M, Gemmel M, Teuwen DE, et al. (12 co-authors). 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 455:661–664.
Worobey M, Telfer P, Souquière S, et al. (14 co-authors). 2010. Island biogeography reveals the deep history of SIV. *Science* 329:1487.
Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol*. 10:1396–1401.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol*. 15:568–573.
Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641–1650.
Zhou Y, Holmes EC. 2007. Bayesian estimates of the evolutionary rate and age of hepatitis B virus. *J Mol Evol*. 65:197–205.
Zlateva KT, Lemey P, Moës E, Vandamme AM, Van Ranst M. 2005. Genetic variability and molecular evolution of the human respiratory syncytial virus subgroup B attachment G protein. *J Virol*. 79:9157–9167.