Evolutionary Origins of Human Herpes Simplex Viruses 1 and 2

Joel O. Wertheim,*[1] Martin D. Smith,[2] Davey M. Smith,[1,3] Konrad Scheffler,[1,4] and Sergei L. Kosakovsky Pond[1]

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

[2]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego

[3]Veterans Affairs San Diego Healthcare System, San Diego, CA

[4]Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa

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

Associate editor: Beth Shapiro

Abstract

Herpesviruses have been infecting and codiverging with their vertebrate hosts for hundreds of millions of years. The primate simplex viruses exemplify this pattern of virus–host codivergence, at a minimum, as far back as the most recent common ancestor of New World monkeys, Old World monkeys, and apes. Humans are the only primate species known to be infected with two distinct herpes simplex viruses: HSV-1 and HSV-2. Human herpes simplex viruses are ubiquitous, with over two-thirds of the human population infected by at least one virus. Here, we investigated whether the additional human simplex virus is the result of ancient viral lineage duplication or cross-species transmission. We found that standard phylogenetic models of nucleotide substitution are inadequate for distinguishing among these competing hypotheses; the extent of synonymous substitutions causes a substantial underestimation of the lengths of some of the branches in the phylogeny, consistent with observations in other viruses (e.g., avian influenza, Ebola, and coronaviruses). To more accurately estimate ancient viral divergence times, we applied a branch-site random effects likelihood model of molecular evolution that allows the strength of natural selection to vary across both the viral phylogeny and the gene alignment. This selection-informed model favored a scenario in which HSV-1 is the result of ancient codivergence and HSV-2 arose from a cross-species transmission event from the ancestor of modern chimpanzees to an extinct *Homo* precursor of modern humans, around 1.6 Ma. These results provide a new framework for understanding human herpes simplex virus evolution and demonstrate the importance of using selection-informed models of sequence evolution when investigating viral origin hypotheses.

Key words: **co-divergence, molecular clock, zoonosis, cross-species transmission, homo, selection.**

Introduction

Herpesviridae is a family of DNA viruses, which epitomize the pattern of viral codivergence with their vertebrate hosts, dating back hundreds of millions of years (McGeoch and Cook 1994; McGeoch et al. 1995). Across the three subfamilies (alpha-, beta-, and gammaherpesvirinae), there have been multitudes of within-host viral lineage duplications (sympatric divergence events [Kitchen et al. 2011]) in which viral descendants follow the phylogenetic history of their host species. Herpes simplex viruses in primates exemplify this phenomenon and evidence of codivergence dates back at least to the ancestor of New World and Old World primates (Simiiformes) 44.2 Ma (Eberle and Black 1993; Luebcke et al. 2006; Steiper and Young 2009) (fig. 1). These viruses have been characterized in monkeys, where each host species is infected with a single, species-specific virus. Humans are the only primate species in which more than one herpes simplex virus has been characterized: HSV-1 and HSV-2 (table 1).

The discovery of the first nonhuman ape simplex virus, chimpanzee herpes simplex virus (ChHV), shed light on how two distinct human simplex viruses arose (Luebcke et al. 2006; Severini et al. 2013). HSV-2 is more closely related to ChHV than it is to HSV-1, which suggests that ChHV and at least one of the human herpes simplex viruses arose via host–virus codivergence. Therefore, there are ten parsimonious scenarios explaining the phylogenetic relationship among HSV-1, HSV-2, and ChHV that require only a single viral lineage duplication or cross-species transmission event (figs. 2 and 3). All of these scenarios predict extant or extinct undiscovered simplex viruses in apes: bonobos (*Pan paniscus*), gorillas, orangutans, and gibbons (shown in gray in figs. 2 and 3). Primate simplex viruses could have undergone within-host viral lineage duplication after apes diverged from Old World monkeys but before humans diverged from the *Pan* genus (fig. 2). Alternatively, HSV-1 could be the result of a cross-species transmission event from gibbons, orangutans, or gorillas (fig. 3*A*–*C*). Finally, HSV-2 could be the result of a cross-species transmission from *Pan troglodytes*, *P. paniscus*, or their common ancestor (fig. 3*D*–*F*).

Although host–virus codivergence is the primary mode of evolution for primate simplex viruses, zoonotic transmission can occur. For example, humans are frequently infected by a macaque simplex virus (MHV-1, formerly known as B virus) (Elmore and Eberle 2008), resulting in severe illness, though human-to-human transmission is exceptionally rare (Centers for Disease Control and Prevention 1987). Moreover, CeHV-2 (previously known as SA8) was discovered in an African green monkey (Malherbe and Harwin 1958), though it has become clear that CeHV-2 is likely a baboon virus (Malherbe and Strickland-Cholmley 1969a, 1969b; Kalter et al. 1978; Hilliard et al. 1989; Tyler and Severini 2006). In the case of human herpes simplex viruses, molecular sequence dating could be used to identify which divergence event (i.e., HSV-1/ChHV or HSV-2/ChHV) corresponds to the speciation between *Homo sapiens* and *P. troglodytes* around 6 Ma (Kumar et al. 2005). Previous dating analysis accompanying the discovery of ChHV used pairwise genetic distance regression analysis to suggest that HSV-2 was the result of codivergence 6 Ma and that HSV-1 originated from an orangutan cross-species transmission event (fig. 3*B*) (Luebcke et al. 2006; Severini et al. 2013). However, such analyses may be misleading because, unlike phylogeny-based molecular clocks, they do not account for shared evolutionary history (Drummond et al. 2003).

Recent methodological developments in the dating of RNA viruses have suggested that standard evolutionary models used in molecular dating can underestimate the time to most recent common ancestor (tMRCA) (e.g., in measles virus, Ebola virus, avian influenza virus, and coronaviruses) (Wertheim and Kosakovsky Pond 2011; Wertheim et al. 2013). This bias has been attributed to the action of strong purifying selection over long evolutionary time scales, and selection-informed models have been shown to improve branch length estimation (Wertheim and Kosakovsky Pond 2011); even under these models, however, many viruses are likely too old to produce reliable tMRCA estimates. For viruses such as HIV and influenza A virus, whose evolutionary rate is on the order of 10^{-3} substitutions/site/year, it is standard to rely on tMRCA estimates around 100 years old (Korber et al. 2000; Worobey et al. 2008, 2014; Smith et al. 2009). The herpes simplex virus substitution rate is estimated to be around 10^{-8} substitutions/site/year (Sakaoka et al. 1994; Norberg et al. 2011), which suggests that tMRCAs on the order of tens of millions of years could be reliably inferred.

To resolve the origin of HSV-1 and HSV-2, we propose a novel computational hypothesis testing approach that

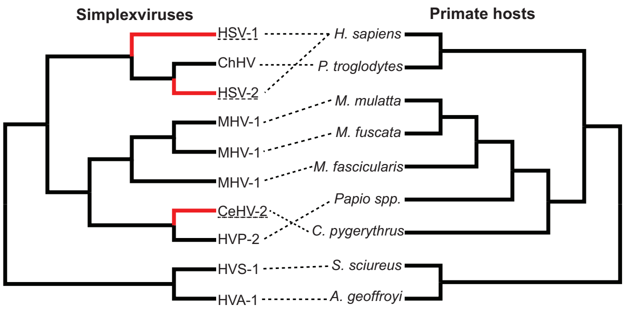

Fig. 1. General pattern of codivergence for primate herpes simplex viruses and their Simiiforme hosts. Underlined viral taxa indicate phylogenetic incongruence, implying cross-species transmission events. Dashed lines connect virus to host species.

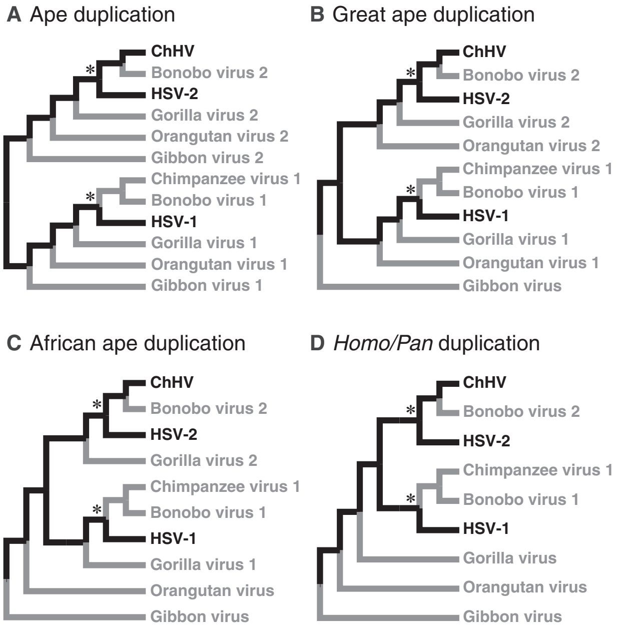

Fig. 2. Evolutionary scenarios that could produce the human and chimpanzee herpes simplex virus phylogeny via viral duplication within a host lineage. Hypothetical unobserved viruses are shown in gray. Nodes representing the common ancestor of humans and chimpanzees, around 6 Ma, are indicated with asterisks. All scenarios imply a 6 Ma tMRCA for HSV-2/ChHV and a tMRCA >6 Ma for HSV-1/ChHV. (*A*) Viral duplication prior to the diversification of apes. (*B*) Viral duplication prior to the diversification of the great apes. (*C*) Viral duplication prior to the diversification of the African apes. (*D*) Viral duplication prior to the split between the *Homo* and *Pan* genera.

Table 1. Known Primate Herpes Simplex Viruses.

Virus	Virus Abbreviation	Host Latin Name	Host Common Name
Baboon herpes virus 2	HVP-2	*Papio* spp.	Baboons
Cercopithecus herpes virus 2	CeHV-2	*Chlorocebus pygerythrus*[a]	African green monkey[a]
Chimpanzee herpes virus	ChHV	*Pan troglodytes*	Chimpanzee
Herpes simplex virus 1	HSV-1	*Homo sapiens*	Human
Herpes simplex virus 2	HSV-2	*H. sapiens*	Human
Macacine herpes virus 1	MHV-1	*Macaca* spp.	Macaques
Saimiriine herpes virus	HVS-1	*Saimiri sciureus*	Squirrel monkey
Spider monkey herpes virus	HVA-1	*Ateles geoffroyi*	Spider monkey

[a] Species is likely not the natural host.

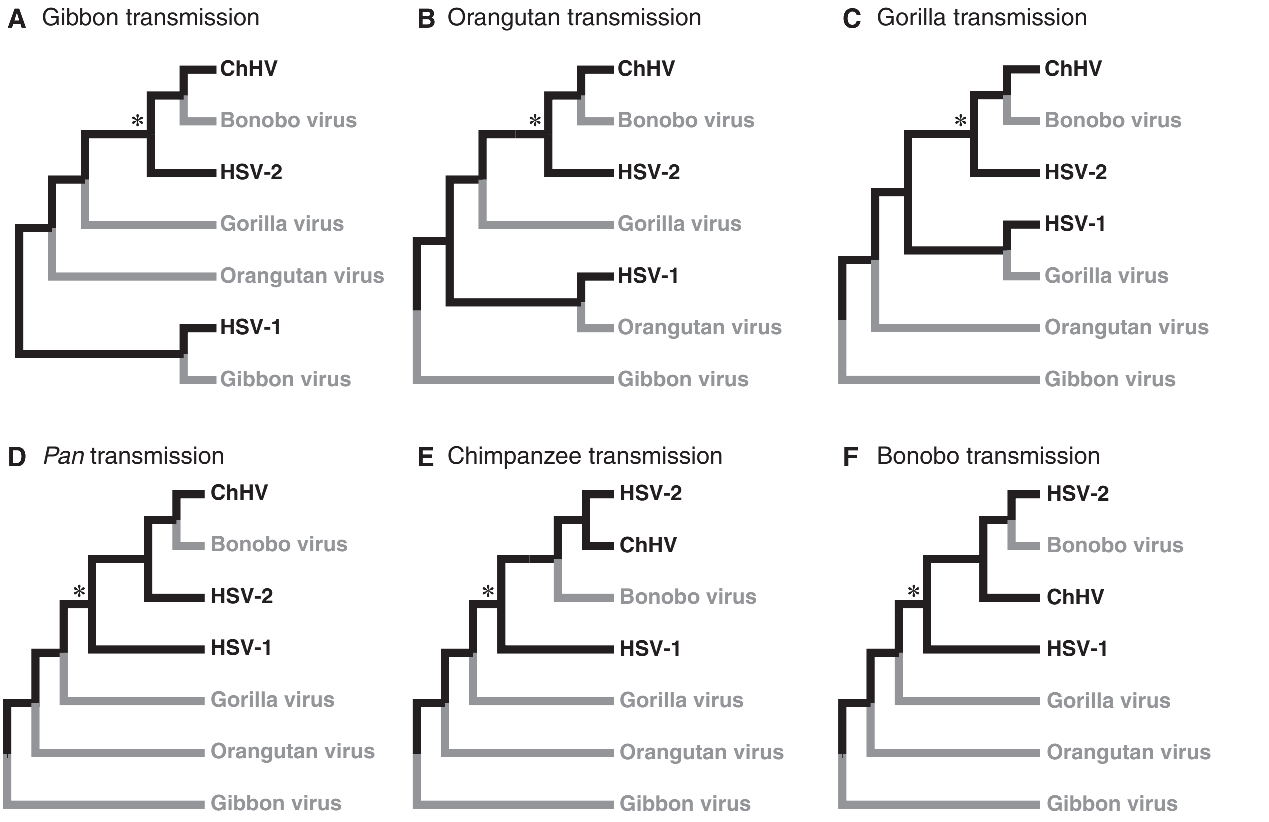

FIG. 3. Evolutionary scenarios that could produce the human and chimpanzee herpes simplex virus phylogeny via viral cross-species transmission. Hypothetical unobserved viruses are shown in gray. Nodes representing the common ancestor of humans and chimpanzees, around 6 Ma, are indicated with asterisks. Scenarios *A*–*C* imply a 6 Ma tMRCA for HSV-2/ChHV and a tMRCA >6 Ma for HSV-1/ChHV, whereas scenarios *D*–*F* imply a 6 Ma tMRCA for HSV-1/ChHV and a tMRCA <6 Ma for HSV-2/ChHV. (*A*) Cross-species transmission of a gibbon virus to a human ancestor, giving rise to HSV-1. (*B*) Cross-species transmission of an orangutan virus to a human ancestor, giving rise to HSV-1. (*C*) Cross-species transmission of a gorilla virus to a human ancestor, giving rise to HSV-1. (*D*) Cross-species transmission of a virus infecting a *Pan* ancestor to a human ancestor, giving rise to HSV-2. (*E*) Cross-species transmission of a chimpanzee virus to a human ancestor, giving rise to HSV-2. (*F*) Cross-species transmission of a bonobo virus to a human ancestor, giving rise to HSV-2.

combines realistic codon-substitution evolutionary models that allow for selection strength heterogeneity with a penalized likelihood molecular clock estimation procedure. The standard evolutionary models produce dating estimates that are not consistent with any of the ten evolutionary origin scenarios. In contrast, our selection-informed model suggests a new explanation for the origin of human herpes simplex viruses, in which HSV-2 was acquired by an extinct *Homo* species from an ancestor of modern chimpanzees.

Results

Bayesian Markov Chain Monte Carlo Dating Analysis

To estimate the tMRCAs of HSV-1, HSV-2, and ChHV, we performed molecular dating analysis on a genome-wide data set comprising 12 concatenated glycoprotein sequences. This analysis was carried out in a Bayesian Markov chain Monte Carlo (BMCMC) framework using a standard nucleotide substitution model (GTR + Γ_4) under a relaxed molecular clock in the BEAST software package (Drummond and Rambaut 2007; Drummond et al. 2012). The molecular clock was calibrated assuming a general pattern of viral–host codivergence across the phylogeny, using three previously estimated internal node ages (see Materials and Methods for details). Under both general sets of hypotheses (viral lineage duplication [fig. 2] or cross-species transmission [fig. 3]), the tMRCA of either HSV-1/ChHV or HSV-2/ChHV should correspond to the divergence between their human and chimpanzee hosts around 6 Ma (fig. 1). We also performed dating analysis using only glycoprotein B (gB) sequences, which have been sampled in a greater number of taxa, permitting the inclusion of an additional calibration point in New World monkeys (see Materials and Methods).

The phylogenetic relationships among the primate simplex viruses were all well supported in the BMCMC analyses in both data sets (posterior probability = 1.0; supplementary fig. S1, Supplementary Material online); the topologies are in agreement with previous viral analyses (Luebcke et al. 2006; Severini et al. 2013) and generally congruent with the host phylogeny (fig. 1). Surprisingly, in the concatenated glycoprotein analysis, neither the HSV-1/ChHV tMRCA (mean = 10.2 Ma, 95% highest posterior density [HPD] = 6.2–14.5 Ma) nor the HSV-2/ChHV tMRCA (mean = 3.2 Ma, 95% HPD = 1.5–5.4) corresponded to the speciation between their human and chimpanzee hosts around 6 Ma (fig. 4*A*). The same pattern recurred in the gB data set: HSV-1/ChHV tMRCA (mean = 10.4 Ma, 95% HPD = 5.3–16.0 Ma) and

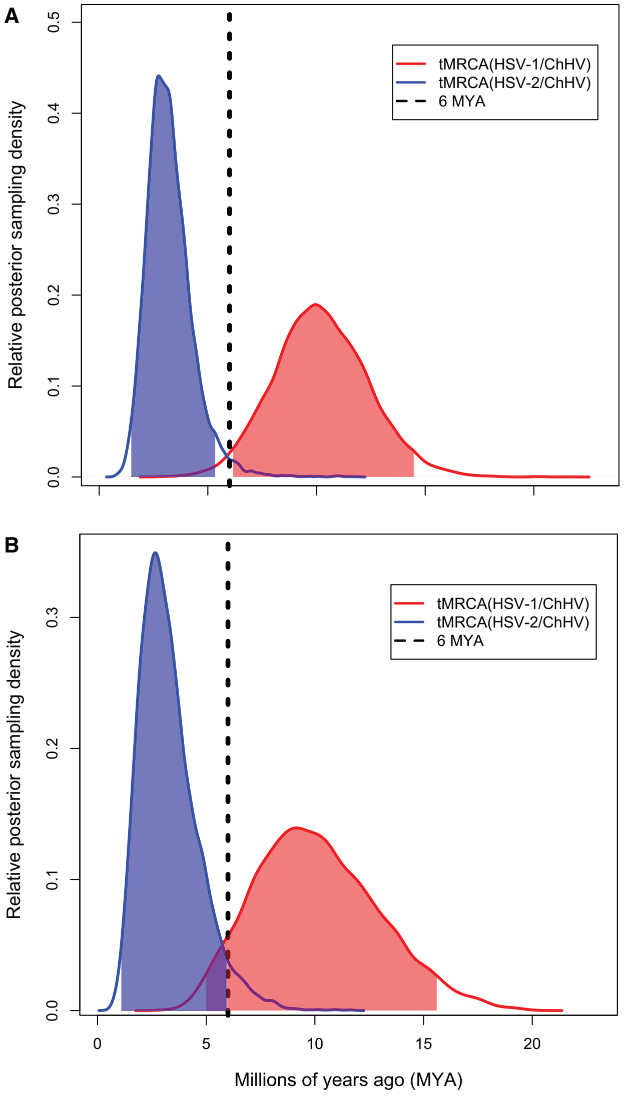

FIG. 4. Posterior distributions of the tMRCAs for HSV-1/ChHV and HSV-2/ChHV in BMCMC analysis. (*A*) Concatenated glycoprotein tMRCA estimates. (*B*) gB tMRCA estimates. Shaded regions depict the 95% highest posterior densities. The vertical dashed line represents the divergence between humans and chimpanzees around 6 Ma.

HSV-2/ChHV tMRCA (mean = 3.4 Ma, 95% HPD = 1.1–5.9) (fig. 4*B*). Although the split between humans and chimpanzees is likely better reflected by a continuous process whereby alleles segregate over a period of time between 5 and 7 Ma (Kumar et al. 2005; Patterson et al. 2006; Yamamichi et al. 2012), BMCMC dating results support neither the lineage duplication nor the cross-species transmission hypotheses for the origin of HSV-1 and HSV-2.

Correcting for Selection-Induced Bias in Branch Length Estimates

We continued with an in-depth analysis of the gB data set, because it contains three additional host species for molecular clock calibration and validation. We inferred a maximum likelihood phylogeny for the primate simplex viruses using this gB data set (fig. 5). Again, the phylogenetic relationships among the primate simplex viruses were well supported (approximate likelihood ratio test [aLRT] = 1.0). Standard nucleotide models (e.g., the ubiquitous GTR + Γ_4) are known to underestimate branch lengths in RNA viruses, resulting in biased tMRCA inference (Wertheim and Kosakovsky Pond 2011; Wertheim et al. 2013). Therefore, we re-estimated the phylogenetic branch lengths using a model of molecular evolution that accounts for variation in selection pressures both across sites and across the phylogeny: branch-site random effects likelihood (BSREL; Kosakovsky Pond et al. 2011).

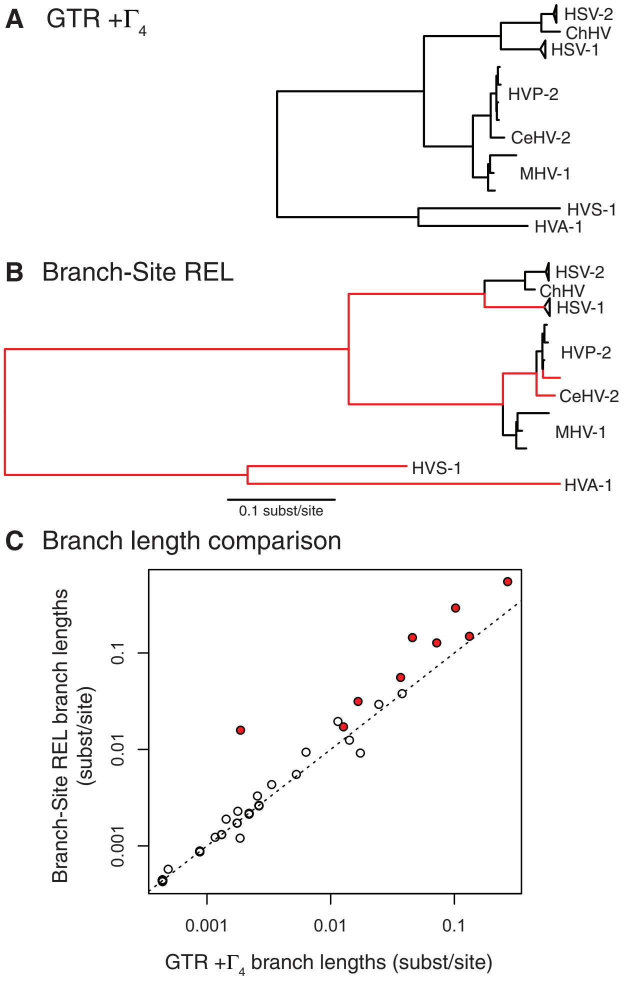

FIG. 5. Branch length expansion under BSREL relative to GTR + Γ_4 substitution model in the gB phylogeny. (*A*) Maximum likelihood tree with branch lengths estimated under GTR + Γ_4. HSV-1 and HSV-2 clades are collapsed. (*B*) Maximum likelihood tree with branch lengths re-estimated under BSREL. Branches determined by cAIC to support multiple d*N*/d*S* classes are colored, and the HSV-1 and HSV-2 clades are collapsed. Both trees are shown on the same scale. (C) Comparison of branches inferred under BSREL and GTR + Γ_4. Branches determined by cAIC to support multiple d*N*/d*S* classes are filled. All other branches supported only a single d*N*/d*S* class. The dashed line depicts *x* = *y*.

A comparison of the branch length estimates under the standard model (GTR + Γ_4) and BSREL shows how the nucleotide model underestimates branch lengths for long branches (fig. 5). Nine branches had statistical evidence of multiple selection regimes (different d*N*/d*S* ratio classes), indicating that a complex pattern of selection has likely acted across the phylogeny. All but two of these branches had weights of 0.95 or greater assigned to the $dN/dS = 0$ class (i.e., the class with no nonsynonymous substitutions). These nine branches tended to be long internal branches and experienced the greatest expansion under BSREL compared with GTR + Γ_4. For instance, the longest internal branch, which separates Old World primate viruses from New World monkey viruses, was 3.15 times longer under BSREL than GTR + Γ_4. These findings suggest that tMRCAs obtained using a standard model are likely biased. One of the length expansions (involving a short terminal branch leading to an HVP-2 isolate) appears to be an outlier (fig. 5C), likely caused by low precision point estimates of dN/dS. However, when we used the original (GTR + Γ_4) branch length instead of the expanded length for this branch in our subsequent analyses, the hypothesis testing results and tMRCA inference remained unaffected (not shown).

Comparison of Evolutionary Scenarios

We inferred the tMRCAs of HSV-1/ChHV and HSV-2/ChHV under the BSREL substitution model and investigated which of the ten alternative hypotheses was most consistent with these estimates. Because codon-substitution models that explicitly include variation in the strength of natural selection cannot be readily implemented in BEAST or similar packages, we employed a penalized likelihood framework (r8s) (Sanderson 2003) to apply a molecular clock using branch lengths previously estimated using BSREL. We calibrated the molecular clock using the same four internal node ages used in the BMCMC analysis.

We constructed two constrained models describing the ten evolutionary scenarios (figs. 2 and 3). The first model, corresponding to viral lineage duplication or HSV-1 cross-species transmission, forces the tMRCA of HSV-2/ChHV to coincide with human–chimpanzee speciation 6 Ma (figs. 2 and 3A–C). The second model, corresponding to HSV-2 cross-species transmission, forces the tMRCA of HSV-1/ChHV to coincide with human–chimpanzee speciation 6 Ma (fig. 3D–F). We then performed likelihood ratio tests comparing the fit of these constrained models to the fit of an unconstrained model in which the tMRCAs of HSV-1/ChHV and HSV-2/ChHV were free to vary. Under the BSREL model, one set of origin scenarios is clearly favored (table 2). A tMRCA for HSV-2 and ChHV at 6 Ma is rejected in favor of an unconstrained model ($P < 0.0001$). Therefore, viral duplication and HSV-1 cross-species transmission scenarios (fig. 2 and 3A–C) can be rejected. However, a tMRCA for HSV-1 and ChHV at 6 Ma (fig. 3D–F) cannot be rejected in favor of an unconstrained model ($P = 0.506$).

Under the GTR + Γ_4 substitution model, a 6 Ma tMRCA for HSV-1/ChHV also provided a better fit than a 6 Ma tMRCA for HSV-2/ChHV (table 2). However, both constrained scenarios were rejected when compared with the unconstrained scenario ($P < 0.0001$). Therefore, if GTR + Γ_4 is an appropriate evolutionary model for the primate herpes simplex viruses, then neither the divergence of HSV-1 nor that of HSV-2 from ChHV corresponded to a host speciation event. Given that prior studies of ancient viral evolution have cast doubt on the suitability of GTR + Γ_4 in this context (Wertheim and Kosakovsky Pond 2011; Wertheim et al. 2013), the problem likely lies with the substitution model and not with the codivergence scenarios.

The divergence between HSV-1 and ChHV could have occurred at any time during the host speciation process, which may or may not have involved substantial gene flow between nascent species (Patterson et al. 2006; Yamamichi et al. 2012). However, the results of our hypothesis tests are qualitatively similar if we allow the divergence between humans and *P. troglodytes* to vary between 5 and 7 Ma (by replacing the fixed constraint with a range); both scenarios are rejected under GTR + Γ_4 ($P < 0.0001$), and only the HSV-2/ChHV codivergence scenario can be rejected under BSREL ($P < 0.0001$).

The maximum likelihood estimate for the HSV-1/ChHV tMRCA using branch lengths inferred under BSREL was 6.2 (5.6–7.4) Ma (table 3). This date is in line with expected divergence between their primate host species. Moreover, the inferred substitution rate of 1×10^{-8} (8.1×10^{-7}–1.2×10^{-8}) substitutions per site per year agrees with previous estimates for herpes simplex virus evolution in humans (Sakaoka et al. 1994; Norberg et al. 2011). The tMRCA of HSV-2/ChHV was 1.6 (1.4–2.1) Ma, suggesting a more recent viral cross-species transmission event. The tMRCAs of HSV-1/ChHV and HSV-2/ChHV are robust to inclusion of specific

Table 2. Maximum Penalized Likelihood Values for gB Phylogenies and Likelihood Ratio Tests for Human and ChHV Codivergence Scenarios under Different Evolutionary Models: GTR + Γ_4 and BSREL.

Homo–Pan Codivergence Event	GTR + Γ_4			BSREL		
	−ln *L*	Δln *L*[a]	*P*[b]	−ln *L*	Δln *L*[a]	*P*[b]
Unconstrained	−307.517	–	–	−628.711	–	–
HSV-1/ChHV	−326.648	19.131	<0.0001	−628.932	0.221	0.506
HSV-2/ChHV	−337.543	30.026	<0.0001	−740.962	112.251	<0.0001

[a] Difference in ln *L* between constrained (null) model and unconstrained (alternative) model.

[b] Likelihood ratio test with 1 degree of freedom comparing the constrained and unconstrained models.

Table 3. Simplex Virus Time of Most Common Ancestor (tMRCA) Estimates from Relaxed Molecular Clock Analyses in r8s and Their Corresponding Host Divergence Dates Inferred Using gB.

Taxa	tMRCA under GTR + Γ_4[a] (Ma)	tMRCA under BSREL[a] (Ma)	Published Host Divergence[b] (Ma)
HSV-1/ChHV	9.1 (8.6–9.5)	6.2 (5.6–7.4)	5–7
HSV-2/ChHV	2.7 (2.9–2.9)	1.6 (1.4–2.1)	5–7
MHV-1	3.7 (3.5–4.1)	2.1 (1.8–2.8)	2.2–2.5
MHV-1 (*Macaca fuscata*/ *M. mullata*)	3.3 (3.1–3.7)	2.0 (1.6–2.5)	1.7
CeHV-2/HVP-2	3.1 (2.9–3.4)	1.9 (1.6–2.5)	11.6

[a] Variance estimates obtained via LHC scheme (see Materials and Methods).

[b] See text for corresponding citations.

internal calibration points. When the penalized likelihood analysis is performed with only three of the four calibrations (all four possible combinations were explored), the tMRCA for HSV-1/ChHV ranges between 5.7 and 6.5 Ma, and the tMRCA for HSV-2/ChHV ranges between 1.5 and 1.7 Ma. Although this test for robustness is potentially sensitive to dependencies that may exist among the nonfossil-derived calibration points, divergence dates within the primate phylogeny have been extensively studied and are internally consistent (Steiper and Young 2009).

tMRCA Inference for Other Primate Simplex Viruses

We also investigated whether BSREL provided consistent tMRCA estimates for the MHV-1. Divergence times within *Macaca* have been well characterized. The three host species whose viruses were included here (i.e., *Macaca mullata*, *M. fuscata*, and *M. fascularis*) share an MRCA around 2.2–2.5 Ma (Tosi et al. 2003). The tMRCA for all three viral lineages estimated using BSREL branch lengths, 2.1 (1.8–2.8) Ma, was closer to the host tMRCA, compared with branch lengths inferred under GTR + Γ_4, which yielded an older tMRCA, 3.7 (3.5–4.1) Ma (table 3). The same pattern holds for the *M. fuscata* and *M. mullata* tMRCA around 1.7 Ma (Fabre et al. 2009). The BSREL analysis estimated a viral tMRCA at 2.0 (1.6–2.5) Ma, whereas the GTR + Γ_4 placed the tMRCA at 3.3 (3.1–3.7) Ma (table 3). Therefore, BSREL provides more internally consistent tMRCAs than the GTR + Γ_4 model and supports a general pattern of codivergence throughout the primate simplex viruses.

An exception to this general pattern of codivergence is found by examining the tMRCA of CeHV-2 and HVP-2. Both substitution models indicate that this tMRCA is too recent to be the result of codivergence with *Chlorocebus pygerythrus* and *Papio* spp. around 11.6 Ma (Raaum et al. 2005) (table 3). The genus *Papio* started diverging around 2 Ma (Sithaldeen et al. 2009; Zinner et al. 2009). Therefore, the CeHV-2/HVP-2 tMRCA inferred with BSREL branch lengths of 1.9 (1.6–2.5) Ma confirms that CeHV-2 is, evolutionarily, a baboon virus, as has been suggested previously (Malherbe and Strickland-Cholmley 1969a, 1969b; Kalter et al. 1978; Hilliard et al. 1989; Tyler and Severini 2006).

Discussion

The evolutionary origins of human herpes simplex viruses can be resolved using phylogenetic and molecular dating analyses. The discovery of ChHV and its placement in the phylogenetic tree yielded several evolutionary scenarios that could explain the origins of HSV-1 and HSV-2. We were able to reject 1) scenarios in which HSV-1 and HSV-2 arose due to viral lineage duplication in apes and 2) scenarios in which HSV-1 is the result of cross-species transmission. Instead ours results suggest 3) a scenario in which HSV-2 is the result of cross-species transmission and HSV-1 is the result of host–virus codivergence. Specifically, the molecular clock analysis indicates that after HSV-1 and ChHV codiverged around 6 Ma, ChHV was transmitted to an ancestor of modern humans around 1.6 Ma, giving rise to HSV-2. Dating estimates within the *Pan* genus provide guidance for distinguishing among the possible HSV-2 origin scenarios (fig. 3*D*–*F*). *Pan troglodytes* diverged from *P. paniscus* around 2.2 Ma and split into four subspecies starting around 1.0 Ma (Stone et al. 2010; Bjork et al. 2011). Therefore, the HSV-2/ChHV tMRCA points to transmission of the precursor of HSV-2 from the common ancestor of *P. troglodytes* to a now extinct *Homo* species (e.g., *H. habilis*, *H. erectus*, and *H. ergaster* [Anton 2003; Severini et al. 2013]) that preceded modern humans (fig. 3*E*). The specific identity of this *Homo* ancestor cannot be determined based solely on the molecular clock. Moreover, recent fossil evidence suggests that these taxa may be best classified as a single *Homo* species (Lordkipanidze et al. 2013). Finally, because both human herpes simplex viruses are transmitted via oral and sexual routes (Brugha et al. 1997; Langenberg et al. 1999), the route of viral transmission between *P. troglodytes* and the extinct *Homo* species (e.g., physical or sexual contact) remains unknown.

Primate herpes simplex viruses have experienced synonymous substitutions to a degree, which causes standard evolutionary models (e.g., GTR + Γ_4) to produce biased tMRCA estimates. However, unlike other viruses such as Ebola, avian influenza, and coronaviruses where the temporal signal has been lost due to levels of sequence evolution which saturate even the selection-aware models, the branch lengths in the primate simplex virus phylogeny can still be estimated reliably if selection-informed models of evolution (e.g., BSREL) are implemented. Our results suggest that primate simplex viruses are young enough to contain sufficient evolutionary signal for molecular dating but too old for this signal to be extracted by standard evolutionary models; BSREL produces internally consistent dating estimates across the primate simplex virus phylogeny. Therefore, selection-informed models should be employed when investigating ancient evolution in both RNA and DNA viruses.

The phylogenetic history of HSV-1, like that of varicella zoster virus (Grose 2012) and *Helicobacter pylori* (Linz et al. 2007), recapitulates human migration patterns dating back tens of thousands of years (Norberg et al. 2011; Kolb et al. 2013). However, molecular clock dating estimates for the tMRCA of HSV-1 may need to be revisited in light of the findings presented here. Specifically, Norberg et al. (2011) inferred a tMRCA for major HSV-1 clades at 710,000 years ago using a calibration based on the tMRCA of HSV-1/HSV-2 at 8.45 Ma, which would correspond to either the African ape duplication (fig. 2*C*) or gorilla cross-species transmission (fig. 3*C*) scenarios. In contrast, Kolb et al. (2013) inferred a tMRCA of major HSV-1 clades around 50,000 years ago and a tMRCA for HSV-1/HSV-2 at 2.2 Ma. Based on this latter age estimate for HSV-1/HSV-2, they postulated a viral duplication event corresponding to the rise of the genus *Homo*; however, the close phylogenetic relationship between HSV-2 and ChHV would necessitate a less parsimonious explanation in which HSV underwent a duplication event followed by cross-species transmission to *P. troglodytes*. More accurate molecular dating within HSV-1 and HSV-2 may be possible using the divergence between humans and *P. troglodytes* as a calibration, though it is possible that standard substitution models may slightly underestimate the evolutionary distance between HSV-1 and HSV-2 (fig. 5*B*).

Our results strongly suggest a scenario (fig. 3*E*) in which African and Asian apes are infected with a single herpes simplex virus (barring genus/species specific extinction or duplication events). Serological evidence suggests that wild mountain gorillas (*Gorilla beringei beringei*) are infected with a herpes simplex virus that is distinct from human herpes simplex viruses (Eberle 1992). The legitimacy of our preferred scenario could be confirmed by the identification of putative gorilla or bonobo herpes simplex viruses. Specifically, if the phylogenetic placement of these putative viruses is found to be in agreement with our predicted scenario, it would serve as a confirmation of the promise held by selection-informed models in the study of viral origins.

This work highlights the need for a method that can simultaneously model variable selection pressures (as in BSREL) and estimate tMRCAs (as in BEAST). Selection-informed models could be incorporated into a Bayesian relaxed molecular clock framework and take advantage of recent computational advances for evaluating codon substitution models (Suchard and Rambaut 2009). Moreover, the appropriate number of d*N*/d*S* classes could be sampled as part of the BMCMC to reflect the level of confidence that exists for multiple classes. If, as this and other studies suggest, selection-informed models are necessary for accurate dating of ancient viral divergence events, then their incorporation into relaxed molecular clock methodology will be an important step towards understanding viral evolutionary history.

Materials and Methods

Data Set

Full-length genomes for primate simplex viruses from six host species were downloaded from GenBank. The 12 conserved glycoproteins spanning the genome (UL1, UL10, UL22, UL27, UL44, UL49A, UL53, US4, US5, US6, US7, and US8) were concatenated and aligned using MUSCLE v2.0, based on translated amino acid sequences (Edgar 2004). Regions containing gaps, indicating ambiguity in homology, were then excised. The final concatenated alignment comprised 7,566 nucleotides. We screened for recombination using GARD (Kosakovsky Pond et al. 2006), which failed to detect any significant breakpoints among the concatenated glycoproteins.

In addition, all available full-length UL27 (gB) genes for primate herpes simplex viruses, representing at least nine host species, were downloaded from GenBank. Alignment was performed using the same protocol, resulting in an alignment of 2,283 nucleotide sites. Identical sequences were replaced with a single representative, because they do not add information for fitting substitution models in the phylogenetic likelihood framework, resulting in a final alignment of 74 sequences (all alignments are available as supplementary material, Supplementary Material online).

BMCMC Molecular Clock Analysis

The tMRCA of HSV-1/ChHV and HSV-2/ChHV was inferred for both the concatenated glycoprotein and gB data sets under an uncorrelated lognormal relaxed molecular clock (Drummond et al. 2006) using a GTR + Γ_4 substitution model in BEAST v1.8.0 (Drummond and Rambaut 2007; Drummond et al. 2012). Four independent chains were run for 25 million generations, sampling every 2,500 generations. The effective sample size for all parameters was greater than 200. These chains were combined using LogCombiner, and the maximum clade credibility trees were summarized using TreeAnnotator. The XML input files are available in supplementary material, Supplementary Material online.

The molecular clock was calibrated assuming a pattern of viral–host codivergence using ages from internal nodes assembled from the literature: 1) tMRCA of Old and New World primates, Simiiformes, at 44.2 Ma (Steiper and Young 2009); 2) tMRCA of Old World monkeys, Catarrhini, at 23 Ma (Raaum et al. 2005); 3) tMRCA of *Macaca* spp. and *Papio* spp. at 9.8 Ma (Raaum et al. 2005), and 4) tMRCA of *Saimiri sciureus* and *Ateles geoffroyi* at 14.4 Ma (Fabre et al. 2009). The last of these nodes was relevant only for the gB analyses, as this gene is the only sequence available for the *A. geoffroyi* simple virus, HVA-1. The tMRCAs of Simiiformes and Catarrhini were estimated based on fossil data, whereas the other two tMRCAs were inferred by previous studies using relaxed molecular clocks. Because the *Macaca*/*Papio* tMRCA was the only date published with associated uncertainty, we placed lognormal prior distributions (mean 0; standard deviation 0.56; similar to a previous study of *P. troglodytes* tMRCAs [Bjork et al. 2011]) at the nodes, offset so that the median value of the distribution corresponded to the tMRCA, to allow for a reasonably degree of uncertainty.

Re-Estimating Branch Lengths

For the gB data set, a maximum likelihood phylogeny was inferred using a GTR + Γ_4 model using a subtree pruning and regrafting algorithm in PhyML 3.0 (Guindon and Gascuel 2003; Guindon et al. 2009), available in SeaView4 (Gouy et al. 2010). Branch support was established using the aLRT (Anisimova and Gascuel 2006). To ensure consistency in later comparisons, branch lengths were reoptimized using HyPhy (Kosakovsky Pond et al. 2005); these branch lengths were indistinguishable from those estimated by PhyML.

To estimate branch lengths under a selection-informed model, we modified the BSREL algorithm. In its original form, BSREL assumed three d*N*/d*S* classes along each branch in the phylogeny, each class representing a proportion of sites evolving with particular d*N*/d*S* value, inferred from the sequence alignment (Kosakovsky Pond et al. 2011). However, this model is generally overparameterized, because short branches rarely contain enough information to support more than one d*N*/d*S* class (Wertheim et al. 2013). Although overparameterization is not a substantial problem when the goal of an analysis is to perform a statistical test for selection (Scheffler et al. 2014), it does become a problem when point estimates of model parameters are used for downstream inference. Therefore, we modified the BSREL model via a step-up parameter selection procedure. Initially, each branch is assigned one d*N*/d*S* class. Then, starting with the longest branch, branch-specific d*N*/d*S* classes are added and retained only if there is an improvement in the small-sample corrected AIC (c-AIC) (fig. 5). Once the optimal number of d*N*/d*S* classes has been inferred, the likelihood model is reoptimized, and branch lengths are estimated.

Penalized Likelihood Molecular Clock Analysis

Inferring tMRCAs on a tree with fixed branch lengths cannot be accomplished using existing relaxed molecular clock packages such as BEAST. Furthermore, forcing an ultrametric tree for dating analysis in the BSREL framework would entail the assumption of a strict molecular clock (which is not realistic) and is not feasible in the current implementation. Therefore, we employed a semiparametric penalized likelihood approach, implemented in r8s (Sanderson 2002, 2003), to smooth the GTR + Γ_4 and BSREL trees under a relaxed molecular clock. Inference in r8s requires trees with fixed branch lengths, allowing us to fit a molecular clock and infer tMRCAs on the gB trees with branch lengths previously optimized under both GTR + Γ_4 and BSREL.

The penalized likelihood algorithm in r8s employs a smoothing parameter, which represents the degree to which the assumption of a strict molecular clock has been relaxed; higher values indicate more relaxation. To estimate these parameters, the same four internal node calibrations were used as in the BEAST dating analysis. These ages were treated as fixed points in r8s, rather than lognormal distributions, because r8s does not perform optimally with narrow calibration windows. Using these node calibrations, r8s estimated the optimal smoothing parameters for the GTR + Γ_4 (smoothing = 3.2) and BSREL (smoothing = 100) trees. The r8s input file is available in supplementary material, Supplementary Material online.

Model Comparison via the Likelihood Ratio Test

Statistical significance was assessed using a likelihood ratio test in which the fixed tMRCA is the null model and the unconstrained tMRCA is the alternative model, with one degree of freedom (as the unconstrained model contains one additional parameter to be estimated). This comparison was performed using gB trees with branch lengths estimated under BSREL and GTR + Γ_4. The four internal calibration points (described above) were used.

Variance Estimates Using Latin Hypercube Sampling

To estimate confidence in our dating estimates, we employed a Latin hypercube (LHC) sampling importance resampling scheme, described in detail previously, to draw 500 samples of scaled trees and estimate length variance (Wertheim and Kosakovsky Pond 2011; Wertheim et al. 2013). Briefly, the sampling distribution of each parameter is approximated by the normal distribution centered on the maximum likelihood estimation and, with variance determined by profile likelihood, discretized into 100,000 bins and used to define the LHC in the parameter space. The likelihood is then evaluated for each of the 100,000 parameter vectors defined by the LHC sampling procedure and resampled using a procedure described previously (Kosakovsky Pond et al. 2010). We took these 500 trees from the LHC analysis (for both GTR + Γ_4 and BSREL) and ran them through r8s. The upper and lower 95% bounds are reported as confidence intervals.

Supplementary Material

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

Acknowledgments

The authors thank Mike Sanderson for offering insight into the r8s program. This work was supported by the University of California Laboratory Fees Research Program (grant number: 12-LR-236617), the National Institutes of Health (grant numbers: DA034978, GM093939), the Bioinformatics and Information Technologies Core of the University of California, San Diego Center for AIDS Research (P30 AI036214), and the Department of Veterans Affairs.

References

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol.* 55: 539–552.

Anton SC. 2003. Natural history of Homo erectus. *Am J Phys Anthropol.* Suppl 37:126–170.

Bjork A, Liu W, Wertheim JO, Hahn BH, Worobey M. 2011. Evolutionary history of chimpanzees inferred from complete mitochondrial genomes. *Mol Biol Evol.* 28:615–623.

Brugha R, Keersmaekers K, Renton A, Meheus A. 1997. Genital herpes infection: a review. *Int J Epidemiol.* 26:698–709.

Centers for Disease Control and Prevention. 1987. Epidemiologic notes and reports B-virus infection in Humans—Pensacola, Florida. *Morb Mortal Wkly Rep.* 39:289–290.

Drummond A, Pybus OG, Rambaut A. 2003. Inference of viral evolutionary rates from molecular sequences. *Adv Parasitol.* 54:331–358.

Drummond AJ, Ho SY, 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, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 29: 1969–1973.

Eberle R. 1992. Evidence for an alpha-herpesvirus indigenous to mountain gorillas. *J Med Primatol.* 21:246–251.

Eberle R, Black D. 1993. Sequence analysis of herpes simplex virus gB gene homologs of two platyrrhine monkey alpha-herpesviruses. *Arch Virol.* 129:167–182.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Elmore D, Eberle R. 2008. Monkey B virus (*Cercopithecine herpesvirus 1*). *Comp Med.* 58:11–21.

Fabre PH, Rodrigues A, Douzery EJ. 2009. Patterns of macroevolution among Primates inferred from a supermatrix of mitochondrial and nuclear DNA. *Mol Phylogenet Evol.* 53:808–825.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol.* 27:221–224.

Grose C. 2012. Pangaea and the out-of-Africa model of varicella-zoster virus evolution and phylogeography. *J Virol.* 86:9558–9565.

Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol Biol.* 537:113–137.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52: 696–704.

Hilliard JK, Black D, Eberle R. 1989. Simian alphaherpesviruses and their relation to the human herpes simplex viruses. *Arch Virol.* 109: 83–102.

Kalter SS, Weiss SA, Heberling RL, Guajardo JE, Smith GC 3rd. 1978. The isolation of herpesvirus from trigeminal ganglia of normal baboons (*Papio cynocephalus*). *Lab Anim Sci.* 28:705–709.

Kitchen A, Shackelton LA, Holmes EC. 2011. Family level phylogenies reveal modes of macroevolution in RNA viruses. *Proc Natl Acad Sci U S A.* 108:238–243.

Kolb AW, Ane C, Brandt CR. 2013. Using HSV-1 genome phylogenetics to track past human migrations. *PLoS One* 8:e76267.

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T. 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science* 288:1789–1796.

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

Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol.* 28:3033–3043.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006. GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22:3096–3098.

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

Kumar S, Filipski A, Swarna V, Walker A, Hedges SB. 2005. Placing confidence limits on the molecular age of the human-chimpanzee divergence. *Proc Natl Acad Sci U S A.* 102:18842–18847.

Langenberg AG, Corey L, Ashley RL, Leong WP, Straus SE. 1999. A prospective study of new infections with herpes simplex virus type 1 and type 2. Chiron HSV Vaccine Study Group. *N Engl J Med.* 341: 1432–1438.

Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW, et al. 2007. An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* 445:915–918.

Lordkipanidze D, Ponce de Leon MS, Margvelashvili A, Rak Y, Rightmire GP, Vekua A, Zollikofer CP. 2013. A complete skull from Dmanisi, Georgia, and the evolutionary biology of early *Homo*. *Science* 342: 326–331.

Luebcke E, Dubovi E, Black D, Ohsawa K, Eberle R. 2006. Isolation and characterization of a chimpanzee alphaherpesvirus. *J Gen Virol.* 87: 11–19.

Malherbe H, Harwin R. 1958. Neurotropic virus in African monkeys. *Lancet* 272:530.

Malherbe H, Strickland-Cholmley M. 1969a. Virus from baboons. *Lancet* 2:1300.

Malherbe H, Strickland-Cholmley M. 1969b. Simian herpesvirus SA8 from a baboon. *Lancet* 2:1427.

McGeoch DJ, Cook S. 1994. Molecular phylogeny of the alphaherpesvirinae subfamily and a proposed evolutionary timescale. *J Mol Biol.* 238:9–22.

McGeoch DJ, Cook S, Dolan A, Jamieson FE, Telford EA. 1995. Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. *J Mol Biol.* 247:443–458.

Norberg P, Tyler S, Severini A, Whitley R, Liljeqvist JA, Bergstrom T. 2011. A genome-wide comparative evolutionary analysis of herpes simplex virus type 1 and varicella zoster virus. *PLoS One* 6:e22527.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441:1103–1108.

Raaum RL, Sterner KN, Noviello CM, Stewart CB, Disotell TR. 2005. Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence. *J Hum Evol.* 48:237–257.

Sakaoka H, Kurita K, Iida Y, Takada S, Umene K, Kim YT, Ren CS, Nahmias AJ. 1994. Quantitative analysis of genomic polymorphism of herpes simplex virus type 1 strains from six countries: studies of molecular evolution and molecular epidemiology of the virus. *J Gen Virol.* 75(Pt 3):513–527.

Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol Biol Evol.* 19:101–109.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302.

Scheffler K, Murrell B, Kosakovsky Pond SL. 2014. On the validity of evolutionary models with site-specific parameters. *PLoS One* 9: e94534.

Severini A, Tyler SD, Peters GA, Black D, Eberle R. 2013. Genome sequence of a chimpanzee herpesvirus and its relation to other primate alphaherpesviruses. *Arch Virol.* 158:1825–1828.

Sithaldeen R, Bishop JM, Ackermann RR. 2009. Mitochondrial DNA analysis reveals Plio-Pleistocene diversification within the chacma baboon. *Mol Phylogenet Evol.* 53:1042–1048.

Smith GJ, Bahl J, Vijaykrishna D, Zhang J, Poon LL, Chen H, Webster RG, Peiris JS, Guan Y. 2009. Dating the emergence of pandemic influenza viruses. *Proc Natl Acad Sci U S A.* 106:11709–11712.

Steiper ME, Young NM. 2009. Primates (Primates). In: Hedges SB, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 482–486.

Stone AC, Battistuzzi FU, Kubatko LS, Perry GH Jr, Trudeau E, Lin H, Kumar S. 2010. More reliable estimates of divergence times in *Pan* using complete mtDNA sequences and accounting for population structure. *Philos Trans R Soc London B Biol Sci.* 365: 3277–3288.

Suchard MA, Rambaut A. 2009. Many-core algorithms for statistical phylogenetics. *Bioinformatics* 25:1370–1376.

Tosi AJ, Disotell TR, Morales JC, Melnick DJ. 2003. *Cercopithecine* Y-chromosome data provide a test of competing morphological evolutionary hypotheses. *Mol Phylogenet Evol.* 27:510–521.

Tyler SD, Severini A. 2006. The complete genome sequence of herpesvirus papio 2 (*Cercopithecine herpesvirus 16*) shows evidence of recombination events among various progenitor herpesviruses. *J Virol.* 80:1214–1221.

Wertheim JO, Chu DK, Peiris JS, Kosakovsky Pond SL, Poon LL. 2013. A case for the ancient origin of coronaviruses. *J Virol.* 87:7039–7045.

Wertheim JO, Kosakovsky Pond SL. 2011. Purifying selection can obscure the ancient age of viral lineages. *Mol Biol Evol.* 28:3355–3365.

Worobey M, Gemmel M, Teuwen DE, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW, et al. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 455:661–664.

Worobey M, Han GZ, Rambaut A. 2014. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature* 508: 254–257.

Yamamichi M, Gojobori J, Innan H. 2012. An autosomal analysis gives no genetic evidence for complex speciation of humans and chimpanzees. *Mol Biol Evol.* 29:145–156.

Zinner D, Groeneveld LF, Keller C, Roos C. 2009. Mitochondrial phylogeography of baboons (*Papio* spp.): indication for introgressive hybridization? *BMC Evol Biol.* 9:83.