MBE Advance Access published April 2, 2010

Using Time-Structured Data to Estimate Evolutionary Rates of Double-Stranded DNA Viruses

Submission: Research Article

Cadhla Firth[1], Andrew Kitchen[1], Beth Shapiro[1], Marc A. Suchard[2,3], Edward C. Holmes[1,4], Andrew Rambaut[4,5]

[1]Department of Biology, The Pennsylvania State University, University Park, PA, USA
[2]Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
[3]Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA, USA
[4]Fogarty International Center, National Institutes of Health, Bethesda, MD, USA
[5]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK

Corresponding Author: Cadhla Firth
Mailing Address: 722 W 168th St. 17th floor, Center for Infection and Immunity,
Columbia University, New York, NY 10032, USA
Telephone: 212-342-9031
Fax: 212-342-9044
Email: cbf2118@psu.edu

Keywords: double-stranded DNA viruses, nucleotide substitution rates, evolution, codivergence, variola virus

Running head: Evolutionary Rates of dsDNA Viruses

© The Author 2010. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org

Abstract

Double-stranded (ds) DNA viruses are often described as evolving through long-term codivergent associations with their hosts, a pattern that is expected to be associated with low rates of nucleotide substitution. However, the hypothesis of codivergence between dsDNA viruses and their hosts has rarely been rigorously tested, even though the vast majority of nucleotide substitution rate estimates for dsDNA viruses are based upon this assumption. It is therefore important to estimate the evolutionary rates of dsDNA viruses independent of the assumption of host-virus codivergence. Here, we explore the use of temporally-structured sequence data within a Bayesian framework to estimate the evolutionary rates for seven human dsDNA viruses, including variola virus (the causative agent of smallpox) and herpes simplex virus-1. Our analyses reveal that while the variola virus genome is likely to evolve at a rate of approximately 1 x 10^{-5} substitutions/site/year, and hence approaching that of many RNA viruses, the evolutionary rates of many other dsDNA viruses remain problematic to estimate. Synthetic data sets were constructed to inform our interpretation of the substitution rates estimated for these dsDNA viruses and the analysis of these demonstrated that given a sequence data set of appropriate length and sampling depth, it is possible to use time-structured analyses to estimate the substitution rates of many dsDNA viruses independently from the assumption of host-virus codivergence. Finally, the discovery that some dsDNA viruses may evolve at rates approaching those of RNA viruses has important implications for our understanding of the long-term evolutionary history and emergence potential of this major group of viruses.

Introduction

The relationships between pathogens and their hosts are highly variable, ranging from newly emergent zoonotic infections such as SARS coronavirus (Peiris et al. 2004), to codivergent associations that span millions of years, like those seen in the papillomaviruses (Bernard et al. 2006). Recently, a concerted effort has been made to unravel the phylogenetic and demographic contexts that have led to this diversity of relationships (Pérez-Losada et al. 2007, Rector et al. 2007, Katzourakis et al. 2009, Smith et al. 2009). In particular, methodological advances have enabled the incorporation of temporal information from time-structured sequence data into strict or relaxed molecular clock models (Drummond et al. 2005, 2006, Drummond and Rambaut 2007), which can then be used to estimate the timing of epidemiologically important events (cross-species transmissions, epidemic outbreaks), as well as historically important ones (the origin of multiple viral subtypes, host-pathogen codivergence). RNA viruses are particularly well-suited to analyses of this type as their rapid replication rate, large population sizes and error-prone polymerase result in large amounts of genetic diversity being generated in measurable amounts of evolutionary time. Sequence analysis programs such as BEAST (Bayesian Evolutionary Analysis by Sampling Trees) are of particular importance in this regard as they utilize the genetic variation present in a sample to simultaneously estimate both demographic and evolutionary parameters in the context of time and space (Drummond et al. 2005, 2006, Drummond and Rambaut 2007, Lemey et al. 2009).

Codivergence with host species over thousands or millions of years has often been invoked as the primary evolutionary mechanism shaping the diversity of many DNA (and some RNA) viruses (Beer et al. 1999, Charrel et al. 1999, Sugimoto et al. 2002, Nemirov et al. 2004). Importantly, inferences about key aspects of viral biology including the rate of evolutionary change in the viral genome and the time-scale of speciation events, have often been made based on the assumption of host-virus codivergence, which automatically places the evolution of these viruses on the same scale as their hosts (Nakao et al. 1997, Hughes and Friedman

2000, Sugimoto et al. 2002, Nishimoto et al. 2006, Krumbholz et al. 2008). However, by incorporating timing information from time-sampled (heterochronous) sequences into inferences about the history of viral populations, it is possible to generate independent estimates of the rate and time-scale of virus evolution, without requiring the strong assumption of codivergence. Critically, analyses of this type have demonstrated that some host-virus systems that were once thought to be examples of codivergence may in fact be the result of much more recent evolutionary associations (Shackelton et al. 2006, Romano et al. 2008, Harkins et al. 2009, Ramsden et al. 2009, Lewis-Rogers and Crandall 2009).

Despite wide-spread interest in using novel statistical models that incorporate heterochronous data to answer a wide range of biological questions, concerns have been raised surrounding biases in evolutionary rate estimates that may be inherent to these methods (Ho et al. 2005, Ho and Larson 2006, Emerson 2007, Ho et al. 2007b, Penny 2005, Navascues and Emerson 2009). For example, a time-dependent relationship has been demonstrated to exist for the molecular clock, such that molecular evolution is accelerated on short time-scales (Ho et al. 2005, Ho and Larson 2006, Ho et al. 2007a,b). Concerns have also been raised regarding the tendency of inference tools to recover substitution rates that are too rapid when population structure in the data is unaccounted for, or when inappropriate calibration points are used (Ho et al. 2008, Navascues and Emerson 2009). Ascribing times to samples and accommodating these in analyses is effectively an assertion that these times span a consequential proportion of the total evolutionary history of the taxa in question, and conditions the analysis on rates that are sufficiently high that this is true. Such analyses have been used extensively to estimate the evolutionary and epidemiological characteristics of a wide range of rapidly evolving RNA and single-stranded (ss) DNA viruses, and appear to return estimates that accord well with the known epidemiology of these pathogens (Shackelton et al. 2005, de Oliveira et al. 2006, Bryant et al. 2007, Rambaut et al. 2008, Firth et al. 2009). However, the utility of these methods in estimating the evolutionary dynamics of double-stranded (ds) DNA viruses, which may evolve far more slowly than RNA viruses, has not yet been investigated. Indeed, the use of heterochronous phylogenetic modeling has resulted in surprisingly high evolutionary rate estimates for JC virus (Shackelton et al. 2006), variola virus (Li et al. 2007) and the bacterium *Neisseria gonorrhoeae* (Pérez-Losada et al. 2007).

In this study, we examine the ability of current inference tools to estimate relatively low evolutionary rates such as those thought to commonly characterize dsDNA viruses. Indeed, Rector et al. (2007) suggest that dsDNA viruses are inappropriate for time-structured analyses because their low mutation rates (~0.003 mutations/genome/replication, Drake 1991, Drake and Hwang 2005, Duffy et al. 2008) will lead to immeasurable levels of genetic change over a given sampling interval. This raises the possibility that the high rates of evolutionary change previously reported for dsDNA viruses arise spuriously, and possibly inevitably, from the models employed. However, if no measurable evolutionary change has occurred within a given sampling period, we would expect any analytical outcomes to exhibit behavior consistent with a lack of temporal structure in the data. Here, we use a variety of human dsDNA virus systems to investigate the ability of heterochronous phylogenetic modeling to (i) accurately estimate the fit of a molecular clock to dsDNA virus data sets with varying sample sizes and distributions, (ii) recover the correct nucleotide substitution rate or reveal that there is a lack of temporal structure in the data, and (iii) place the time to the most recent common ancestor (TMRCA) of these viruses within the correct time frame (when known). To further assess the behavior of these analytical tools, we created synthetic data sets under a variety of sampling schemes and substitution rates, from an 'RNA virus-like' rate of $1\text{x}10^{-4}$ subs/site/year to a 'dsDNA virus-like' rate of $1\text{x}10^{-8}$ subs/site/year. Using these data we examined the ability of standard modeling tools to recover the nucleotide substitution rate and TMRCA from a population when the true evolutionary history of a sample is known.

We considered seven dsDNA viruses in our analysis. Human papillomavirus type-16 (HPV-16) is one of more than 100 human viruses within the large *Papillomaviridae* family of dsDNA viruses with small genomes ranging from ~6 to 8 kbp. Many types of HPV (including HPV-16) apparently exhibit strong patterns of codivergence with human populations, mirroring the global movement of humans out of Africa, and clustering phylogenetically by ethnicity rather than by current geographic distribution (Ong et al. 1993, Calleja-Macias et al. 2005, Bernard et al. 2006, Chen et al. 2009). The rate of nucleotide substitution for HPV-18 has been estimated at ~$4.5x10^{-7}$ substitutions per site per year (subs/site/year) based on codivergence with human populations, similar to estimates obtained for the feline papillomaviruses ($1.95x10^{-8}$ subs/site/year), again assuming codivergence (Ong et al. 1993, Rector et al. 2007).

Two members of the *Alphaherpesvirinae* were also included in this analysis: Herpes Simplex Virus-1 (HSV-1), and Varicella-Zoster virus (VZ). Herpesviruses are large dsDNA viruses (genomes range from 125 to 240 kbp) that infect both vertebrates and invertebrates. Phylogenies of the alphaherpesviruses show topologies highly congruent with those of their diverse tetrapod hosts, and evolutionary rates for HSV-1 have been estimated at $3.5x10^{-8}$ to $3.0x10^{-9}$ subs/site/year based on the assumption of a codivergent history with their hosts (Sakaoka et al. 1994, McGeoch et al. 2000). In contrast, the evolutionary history of VZ is less conclusive, with evidence of both a codivergence relationship with humans (Wagenaar et al. 2003) and a more recent origin suggested by phylogenetic analyses (Muir et al. 2002).

BK virus is a human polyomavirus with a ~5 kbp genome that is closely related to JC virus. Polyomaviruses were historically considered examples of human-virus codivergence, and both BK and JC viruses have been used as markers for patterns of human evolution and migration. However, more rigorous phylogenetic analyses of the relationship between a variety of polyomaviruses and their primate hosts have suggested that no significant similarities in tree topology or evolutionary time-scale exist between these groups (Pérez-Losada et al. 2006, Shackelton et al. 2006, Zheng et al. 2007, Krumbholz et al. 2008). Estimates of the rate of evolution of BK virus are greatly affected by the use of different calibration assumptions, ranging from an intra-host rate estimate of $2\text{-}5x10^{-5}$ subs/site/year (Chen et al. 2004), to rates of

1.41×10^{-7} to 4×10^{-8} subs/site/year based on the assumption of codivergence between viral and human populations (Yasunaga and Miyata 1982, Krumbholz et al. 2008).

Variola virus (VARV) is the etiological agent of the human-specific pathogen smallpox, from the *Poxviridae* family (genome size = ~190 kbp). The first unequivocal description of smallpox in human populations occurred in 4th Century A.D. in China, although cases have been suspected as far back as 1122 B.C (Li et al. 2007). Previous estimates based on time-structured sequence data have placed the origin of smallpox at 207-231 ybp using strict and relaxed clocks, respectively (Li et al. 2007). However, this extremely recent estimate has been disregarded as being at odds with both historical/epidemiological data and with the low genetic diversity identified in serially-sampled sequences (Esposito et al. 2006, Li et al. 2007). As a result, these authors have suggested that their evolutionary rate estimates may be upwardly biased by the use of heterochronous phylogenetic models. In contrast, calibration with historical records of smallpox infection (all of which are debatable) placed the TMRCA for VARV at 1,400 - 6,300 ybp, depending on the choice of calibration point, with correspondingly lower substitution rates (Li et al. 2007, Babkin and Shchelkunov 2008, Hughes et al. 2009, Shchelkunov 2009). Importantly, any calibration of contemporary strains using historical records is potentially problematic as selective sweeps and population bottlenecks can purge genetic diversity from the population, resulting in a TMRCA that is far more recent than the historical association of the virus with its host.

The origin(s) and emergence patterns of human adenoviruses (HAdV, genome sizes range from ~26 to 45 kbp) are considerably more vague. The *Adenoviridae* form five distinct clades, corresponding to their mammal, reptile, bird, amphibian and fish hosts. This phylogenetic structure has led to the hypothesis that the five lineages codiverged along with the host classes, with an estimated date for the split between human and chimpanzee adenoviruses at ~5.5 mya (Benkö and Harrach 2003). Subsequent work has demonstrated that while some HAdV groups may show cursory support for a codivergent history with primates (Subtype C), other groups do not (Subtypes B and E). In addition, examination of the primate adenoviruses shows that the majority of human and non-human adenoviruses are mixed throughout the tree (Madisch et al. 2005, Roy et al. 2009).

Materials & Methods

Virus Data Sets

Data sets for each of the seven dsDNA viruses were compiled based on availability in GenBank. Full genome sequences were used when available; otherwise, appropriate smaller gene segments were used to maximize the size of the data sets. In all cases, the sampling year was also collected for each sequence, and only samples that had not been exposed to extensive laboratory manipulation were included. The details of all data sets are shown in table 1. Each data set was aligned manually using Se-Al (v2.0a11 Carbon, http://tree.bio.ed.ac.uk/software/seal) and examined for evidence of recombination using the Bootscan, Chimaera, GENECONV, MaxChi, RDP, and SisScan methods with default parameters, implemented in the RDP3 software package (Martin et al. 2005). Potential recombinant sequences were identified when three or more methods within RDP3 were in agreement with P<0.001. All potential recombinants were removed from further analysis.

Phylogenetic Analysis

Maximum likelihood (ML) phylogenies were estimated using PAUP* (4.0b, Swofford 2003) for each alignment using the tree-bisection-reconnection method of branch swapping and the best nucleotide substitution model as determined by Modeltest (v3.7, Posada and Crandall 1998). The clock-like behavior of each data set was then assessed using a regression of root-to-tip genetic distances inferred from the ML trees against sampling time in the program Path-O-Gen (v1.1, http://tree.bio.ed.ac.uk/software/pathogen/; Drummond et al. 2003). Under this analysis the correlation coefficient indicates the amount of variation in genetic distance that is explained by sampling time. This provides a correlative measure of the goodness-of-fit of the data to a strict molecular clock, and designates a root for the phylogeny that is most consistent with a molecular clock.

Phylogenies incorporating sampling time were then estimated for each data set using the Bayesian Markov Chain Monte Carlo (MCMC) inference methods made available in BEAST (v.1.4.8, Drummond and Rambaut 2007). These analyses were run using either the HKY or GTR model of nucleotide substitution, and with or without an among-site rate heterogeneity parameter (gamma) depending on the model that best fit the data. For protein coding genes, the alignment was also partitioned into codon positions, while the full genome alignments also included a parameter for invariant sites. Genealogies were estimated under (i) a strict molecular clock, (ii) a relaxed molecular clock with an uncorrelated lognormal distribution (UCLN) of rates, and (iii) a relaxed molecular clock with an uncorrelated exponential distribution (UCED) of rates. A variety of prior probability distributions on the parameter characterizing the rate of evolutionary change under a molecular clock were explored. These included both a uniform distribution with boundaries at 0 and 100, and an exponential distribution with mean expectations that ranged from $1.0x10^{-6}$ to 1.0. The evolutionary and coalescent parameters were estimated under both the hypothesis of a constant population size, and using the less constrained Bayesian skyline coalescent model (Pybus et al. 2000, Strimmer and Pybus 2001, Drummond et al. 2005). A minimum of two independent MCMC simulations for each model-clock combination were performed for no less than 100 million generations, sub-sampling every 10,000 generations to decrease auto-correlation between model parameter samples. The two runs were combined for inspection after removing a 10% burnin from each, and statistical confidence in the parameter estimates was assessed by reporting marginal posterior parameter means and their associated 95% highest probability density (95% HPD) intervals.

To test the temporal signature present in these data sets, we used a tip-date randomization technique (Duffy and Holmes 2009, Ramsden et al. 2009). Here, a null distribution of mean substitution rates was generated by randomizing the sampling date associated with each sequence, 20 times per alignment. The substitution rate was then re-estimated in BEAST for each randomized data set under the best-fitting model for the true data, as above. To assess the significance of the temporal structure present in the data, the mean evolutionary rate estimate from the observed data was compared to the 95% HPDs estimated from the randomized data sets. We also expect the lower tail of the 95% HPD of the evolutionary rates from the randomized data to be large, and tend strongly towards zero.

Synthetic Data Sets

Synthetic data sets were used to test the ability of heterochronous phylogenetic modeling to correctly recover the true nucleotide substitution rates from a range of values that may exist in DNA viruses (1.x10^{-4} to 1x10^{-8} subs/site/year), given the type of data used in the first part of this study. Synthetic data sets were generated to reflect the evolutionary processes in two of the seven dsDNA data sets analyzed, VARV and HSV-1, which represent the typical range one might encounter in an analysis of temporally sampled data (table 1). The VARV-like synthetic data consisted of 50 sequences of 100,000 bp sampled from 1946 to 1982 (similar to a full-genome alignment), while the HSV-1-like synthetic data consisted of 84 sequences that were 1200 bp in length, sampled from 1981 to 2008 (a typical single-gene alignment). Sequences for each data set were generated in the following manner. The year-of-sampling distribution associated with the sequences was fixed to those values from the VARV and HSV-1 data sets and random phylogenies were drawn from a coalescent process based on each group of taxa (N = 50 or 84) using an MCMC algorithm and assuming a constant population size. The root height for each phylogeny was fixed at one of the following intervals: 100 ybp, 1000 ybp, 10,000 ybp, 100,000 ybp, and 1,000,000 ybp. Sequences of the appropriate length were simulated along each random tree following the HKY model of sequence evolution with data set-specific base frequencies and transition/transversion ratios based on the ML estimate of these values from the actual virus data sets. Gamma distributed rate variation was also initially added to the HSV-1-like simulations, but as the inclusion of this parameter did not impact the results of the simulations, it was removed from additional replications (data not shown). For the first set of synthetic data, the rate of evolution followed a strict molecular clock (standard deviation around the mean rate of 0.0) set at one of the following values corresponding to each specified value of the root height: 1x10^{-4} subs/site/year (when the root height was 100 ybp), 1x10^{-5} subs/site/year, 1x10^{-6} subs/site/year, 1x10^{-7} subs/site/year and 1x10^{-8} subs/site/year (when the root height was one million ybp). Twenty independent replicate data sets were created for each root height/clock rate, and each virus-type. These synthetic data sets contained a similar number of variant sites to the actual data sets, indicating that we were correctly modeling the true pattern of evolution (data not shown). By co-varying the tree height and substitution rate in this manner, the observed genetic diversity of each synthetic data set was held constant and similar to that observed in the VARV and HSV-1 data sets. This allowed for a direct assessment of the power of this method to estimate substitution rates under successive conditions in which smaller proportions of overall genetic diversity can be attributed to the sampling interval. Posterior inference was performed on each data set assuming a strict molecular clock, the HKY model of nucleotide substitution and constant population size. MCMC simulations were run for 100 million generations, with sub-sampling every 10,000 generations. Convergence of all parameters was verified visually using the program Tracer (http://tree.bio.ed.ac.uk/software/tracer/).

An additional set of simulations was performed using both the HSV-1 and VARV-like data, but incorporating branch-rate heterogeneity into the sequence simulation process. In this case, sequences were evolved assuming a UCLN relaxed molecular clock with mean rates of 1x10^{-6}, 1x10^{-7} and 1x10^{-8} subs/site/year and a standard deviation of 2.0 around the mean, with corresponding root heights as above. Twenty replicate data sets were created with each substitution rate for each virus type, and the posterior inferences were performed assuming a strict clock, as above. The recovered estimates of the nucleotide substitution rates and TMRCA (posterior mean and 95% HPD) from each group of simulations were then compared to the known values of the simulated data. The number of simulations that included the true rate and TMRCA estimates within the 95% HPDs was recorded for each clock rate/root height. In total, 320 MCMC analyses of simulated data were performed, 160 each for VARV- and HSV-1-like data. A sample xml file for use in BEAST is included as Supplementary Information.

Results

Inference of Substitution Rates

The Bayesian MCMC inference of each dsDNA virus data set converged efficiently on a posterior mean value for all parameters with correspondingly narrow posterior distributions around the mean rate and TMRCA. No substantial differences in the posterior estimates of the mean evolutionary rates were observed when the initial values or prior distribution of the rate parameter were varied. The mean evolutionary rates estimated for the seven dsDNA virus data sets ranged from 10^{-6} to 10^{-3} subs/site/year (VARV/VZ and HPV-16, respectively) (table 2). The 95% HPDs ranged from 10^{-7} (HAdV-C) to 10^{-2} subs/site/year (HPV-16), and varied by no more than an order of magnitude from their corresponding posterior mean in either direction (table 2). Interestingly, with the exception of HPV-16, the mean evolutionary rates estimated were also highly consistent across data sets, ranging from 6×10^{-6} to 8×10^{-5} subs/site/year. These estimates were highly robust to choice of clock, rate distribution or demographic model parameters, but were much higher than expected based on previous estimates of the substitution rate of dsDNA viruses (table 2) (Duffy et al. 2008, Bernard 1994). In fact, the evolutionary rates estimated for HPV-16, HSV-1, BK virus, HAdV-B and HAdV-C all approached those measured in RNA viruses (Jenkins et al. 2002, Duffy et al. 2008). In addition, these high rates were associated with extremely recent TMRCA estimates in all cases, ranging from ~10 ybp (HPV-16) to ~800 ybp (BK) (table 2).

Most dsDNA viruses are thought to evolve primarily through codivergence with their hosts, and this process should be reflected in low rates of evolutionary change (Villarreal et al. 2000, Holmes 2004, Holmes and Drummond 2007). If we allow that the seven dsDNA viruses used in this study do indeed evolve at a low rate consistent with a codivergent history (an assumption that is not contentious in many cases), the possibility must be considered that the high substitution rates estimated here are the result of estimation error, and do not reflect the true evolutionary history of the data. As it is possible that such erroneous results are due to a shallow sampling interval relative to the actual time scale of evolution (figure 1), we assessed the strength of the temporal structure in our data sets by utilizing a randomization procedure.

Randomization Test

The sequence-sampling time associations in each data set were randomized 20 times per virus and the evolutionary rates were re-estimated from each of these data sets. The hypothesis of significant temporal structure was rejected when the value of the mean evolutionary rate estimated from the real data fell within the 95% HPDs of those estimated from the randomized data. Comparisons between the actual and randomized estimates for each virus revealed significant support for the presence of temporal structure in the data in all but two cases (HAdV-C and VZ) (figure 2). Indeed, the 95% HPDs of the actual rate estimates fell outside the entire distribution of the randomized data sets for HSV-1, BK, VARV and HAdV-B (figure 2).

Based on this analysis, it is clear that there is no support for a high substitution rate in either HAdV-C or VZ. In contrast, the accuracy of the evolutionary rates for the remaining five data sets cannot be discounted by this measure (figure 2). This latter result is surprising, given the extremely rapid rate estimates that were recovered (table 2). If these sequences were sampled over a time interval that was too narrow relative to the rate of viral evolution (i.e. recent samples of slowly evolving/codiverging viruses), the dates at which the samples were taken should confer little information about the evolutionary dynamics of the virus (figure 1).

Estimates of evolutionary rates from the actual and randomized data sets should then recover comparable values, producing a pattern similar to that seen in HAdV-C and VZ (figure 2). One caveat of this approach stems from the ability of the randomizations to break-up both temporal and geographic structure in the data. Therefore, if hidden population substructure is contributing to erroneous rate estimates (Navascues and Emerson 2009), it would be broken-up by the temporal randomization process.

Root-to-Tip Regressions

A conservative assessment of the degree of clock-like evolution present in a data set is achieved by fitting a regression of the year-of-sampling against the root-to-tip genetic distance of each sample, measured from an ML tree. When this regression was calculated for the seven dsDNA virus data sets studied here, the resultant correlation strongly supported the presence of molecular clock-like structure in VARV, weakly suggested the presence of clock-like evolution in HAdV-B, HSV-1 and VZ, and revealed no support at all for this hypothesis in the HPV-16, BK and HAdV-C data sets (figure 3). These results were not surprising for the HPV-16 data set considering that the sequences were sampled over only a four-year interval (table 1), and are consistent with the lack of temporal structure in the HAdV-C data identified by the randomization analysis (figure 2).

Taken together, the heterochronous phylogenetic modeling analyses, randomization procedure and regression analyses suggest the presence of high evolutionary rates in the HSV-1, BK, VARV, and HAdV-B data sets, with varying levels of consistency. The VARV data set was unique in that the relatively high rates of nucleotide substitution estimated by BEAST (~9.32×10^{-6} subs/site/year, table 2) for this virus were strongly supported by both the randomization and regression analyses (figures 2,3). In contrast, high substitution rates for the HSV-1, BK, and HAdV-B data sets were supported by the randomization procedure, but showed a weaker correlation between root-to-tip genetic distance and sampling time (R^2 = 0.141, 0.004 and 0.327, respectively) than that found in VARV (R^2 = 0.679).

Synthetic Data

When sequences were generated under a model approximating a strict molecular clock (i.e. all branches of the tree evolve at the same rate), posterior inference was able to recover the correct substitution rate with narrow 95% HPDs for all 20 replicates of both the VARV- and HSV-1-like data sets when the rate was 10^{-4} subs/site/year (figure 4). The true substitution rates were returned with similar accuracy and precision for the VARV-like data sets when the rate was 10^{-5} or 10^{-6} subs/site/year (figure 4). However, at 10^{-7} and 10^{-8} subs/site/year the mean substitution rates were consistently higher than the true values, but characterized by long-tail posterior distributions tending towards zero (figure 4). For all VARV-like simulations, the true substitution rate was contained within the 95% HPDs of the estimated rate, as were the estimated TMRCAs (data not shown). We were considerably less successful at recovering the true substitution rates for the HSV-1-like synthetic data sets. When the true substitution rate was 10^{-5} subs/site/year, the posterior mean estimates were close to the true value; however, some of the individual posterior distributions of the rate were highly skewed towards zero (figure 4). Similarly, when the set rate was 10^{-6} subs/site/year or lower, our tools were unable to recover a mean rate that was close to the true value, and the 95% HPDs ranged from the highest possible rate supported by the data, to a value approaching zero (figure 4). We consider these widely skewed posterior distributions of the rate to signify a lack of significant temporal structure in the data, an effect not seen in estimates based on our real data where values close to zero were not observed.

To determine if the high substitution rates recovered for the dsDNA viruses analyzed in the first part of this study could be a result of deviations from the molecular clock model coupled with low temporal signal in the data, we added branch-rate heterogeneity (i.e. relaxed clock behavior) to the synthetic data sets when the mean rate was set to 10^{-6}, 10^{-7} and 10^{-8} subs/site/year, and re-estimated the rate of substitution. The posterior mean and 95% HPDs estimated from these synthetic data sets were similar to those returned from the data simulated under a strict clock (figure 5). When the true rate of the VARV-like data was set at 10^{-6} subs/site/year, the resultant estimates were close to the true values; however, substantial deviations from the known values occurred when the rates were set to 10^{-7} or 10^{-8} subs/site/year, with correspondingly larger 95% HPDs (figure 5). The rates estimated from the HSV-1-like data simulated with branch-rate heterogeneity also closely mirrored those from the data simulated under a strict clock. The mean substitution rates estimated from all HSV-1-like data were higher than the true values, and again associated with wide, long-tail posterior distributions that tended towards zero (figure 5). As before, we consider these distributions to indicate a lack of temporal structure in the data at these low evolutionary rates.

Discussion

Based on the distribution of rates from our synthetic data sets, we are able to make a number of general conclusions about the use of heterochronous data to estimate the substitution rates and divergence times of potentially slowly evolving dsDNA viruses. In particular, given a data set containing a large enough number of variable sites (such as the VARV data set), it is possible to accurately estimate substitution rates that range from 10^{-4} to approximately 10^{-7} subs/site/year using temporally sampled viruses, even if the data do not conform to a strict molecular clock (figures 4, 5). This is dependent on the length of sequence, and for a data set containing only a small number of informative nucleotide sites (as in the case of HSV-1), the temporal signal in the data begins to break down at evolutionary rates below 1×10^{-5} subs/site/year (figures 4, 5). In these cases, the sampling interval is probably too small relative to the rate of evolution of the virus. Therefore, any substitution rate estimated using these types of data will likely not converge on the true rate, but will instead return a wide posterior distribution of the rate that tends towards zero (figure 5). This behavior is also likely to be robust to a poor fit to the molecular clock (e.g. the use of a strict clock with highly rate-variable data), and can be interpreted as an indication that the data are not appropriate for a temporal analysis by a program such as BEAST.

Interestingly, the shape and lower tails of the posterior distributions estimated from our synthetic data at low rates were similar to those returned from analyzing the HPV-16, HAdV-C and VZ data sets (table 2). This strongly suggests that there is insufficient temporal structure in these data to undertake an analysis of the nucleotide substitution rate in the absence of external calibration points. However, a very different pattern was seen in the HSV-1, BK, VARV and HAdV-B data sets (figure 4). Here, analysis of the observed data sets resulted in high substitution rate estimates associated with tight 95% HPDs, a pattern that we were unable to reproduce using the synthetic data even with the inclusion of large amounts of branch-rate heterogeneity (i.e. when applying a standard deviation of $2.0\log_{10}$ to the mean clock rate). As such, extensive rate heterogeneity across the phylogeny is unlikely to explain the high substitution rates observed in these viruses, and we are therefore unable to justify the high rates of evolution measured for some of these dsDNA viruses by simply invoking estimation error. The analysis of contemporaneously sampled, slowly evolving viruses results in a posterior distribution that reveals the lack of structure in the data, a phenomenon unlike the posterior distributions recovered from the analysis of these viruses. Hence, our simulations support the conclusion that VARV is evolving at a rate close to $1\text{x}10^{-5}$ subs/site/year, and indicates that there is no obvious flaw in the analysis of the HSV-1, BK and HAdV-B data sets that could result in erroneous substitution rate estimates. Therefore, we accept high rates of evolution in VARV, and tentatively suggest that high substitution rates may also occur in the thymidine kinase gene of HSV-1, the HAdV-B hexon (capsid) gene, and potentially in BK virus, although further work is clearly needed. It is also possible that these high rates are the result of model misspecification(s) that have not yet been identified. We now consider the case of some of these viruses in more detail.

Variola Virus

The rate of nucleotide substitution in VARV obtained from our analysis (~1x10^{-5} subs/site/year), accords well with rates previously estimated (and discarded) using a variety of methods. Multiple attempts have been made to use historical documentation of the (potentially) first smallpox outbreaks in endemically infected regions as calibration points to estimate the divergence dates and substitution rate of VARV (Li et al. 2007, Babkin and Shchelkunov 2008, Shchelkunov 2009). Accordingly, the choice of calibration points has dramatically influenced many of these estimates, resulting in an approximate TMRCA for VARV ranging from 200 to 6000 ybp, with correspondingly variable substitution rates (Li et al. 2007, Babkin and Shchelkunov 2008, Shchelkunov 2009). In most cases, these externally-calibrated estimates are orders of magnitude lower from those achieved using temporally sampled sequences (Li et al. 2007, this study). In an attempt to clarify the timescale of VARV evolution, Hughes et al. (2009) estimated the substitution rate using the synonymous sites of 132 protein-coding genes distributed throughout the genome, and scaled the rate around the two dates previously suggested for the introduction of the P-II clade into South America (the 16th Century African slave trade or the 18th Century West African slave trade). The use of either of these calibration points had little impact on the overall rate estimate, which ranged from 4 - 6x10^{-6} subs/site/year and was similar to both our rate estimate and that of Babkin and Shchelkunov (2008) (2 - 3x10^{-6} subs/site/year).

Two primary pieces of information have been used to support the idea that VARV is a slowly evolving virus with a long history of association with human populations. The first is that epidemiologically linked isolates appear to accumulate no mutational changes over the period of one year (Li et al. 2007). However, it is also theoretically possible to explain this observation by invoking selective sweeps or severe population bottlenecks during transmission, which act to reduce the variability accrued during replication in a single individual. The second factor opposing high evolutionary rates for VARV is the suggestion of congruence between historical records of the introduction and spread of the virus in various locations, and the TMRCAs estimated for the major VARV clades. However, as noted above, the use of historical records as calibration points from which the rate and divergence times of viruses are estimated is problematic. In particular, these epidemiological records do not necessarily correlate with the true origin or introduction of a pathogen, as revealed by the different dates used for the same introductions of VARV in multiple locations, and the last common ancestor of any contemporaneously sampled set of viruses may be of more recent origin than the common ancestor of the entire species (Hughes et al. 2009, Shchelkunov 2009). Calibrating evolutionary analyses using a time point that predates the age of the most recent common ancestor of a sample will bias the substitution rate and divergence date estimates towards a much older history (and correspondingly lower substitution rates). In contrast, the current analysis only estimates the TMRCA of the sample, not of VARV itself, and may therefore return TMRCA estimates more recent than the known age of this human pathogen, particularly if there has been a large-scale reduction in diversity (i.e. selective sweep or population bottleneck) in the recent history of VARV. Although we cannot exclude all estimation errors in the case of VARV, the high level of genetic variation in our data explained by temporal sampling (R^2 = 0.68) lends support to the existence of rapid evolutionary rates in VARV (Hughes et al. 2009, this study).

Herpes Simplex Virus-1

The cases of HSV-1 and HAdV-B are considerably more perplexing than that of VARV. The estimated substitution rates for these viruses were highly consistent between models (table 2) and were supported by the randomization test of temporal structure in the data (figure 2). In addition, we were unable to reproduce similarly high estimates of the substitution rate from viral data sets known to be slowly evolving through the analysis of synthetic data. In particular, the possibility that genes within herpesviruses such as HSV-1 could be evolving at a rate of 10^{-5} subs/site/year is perhaps one of the most difficult results of this study to reconcile with the biology of the virus as it is presently understood.

Herpesviruses are prototypic examples of host-pathogen codivergence, and comparing the phylogeny of their vertebrate hosts with that of the virus indeed reveals strong topological congruence (McGeoch and Cook 1994, McGeoch et al. 1995, 2000, Jackson 2005). However, the substitution rate for herpesviruses such as HSV-1 has not previously been measured independently from the hypothesis of codivergence. A key measure of the evolutionary rate of herpesviruses originated from the analysis of 63 variable restriction endonuclease sites in 242 HSV-1 samples from three human populations represented by six countries: Asian (Korea, Japan and China), African (Kenya) and European (Sweden and the USA) (Sakaoka et al. 1994). Based on rough estimates of the divergence times between these ethnic groups (110,000 years for the split between African and European/Asian populations, and 50,000 years for the split of Asian from European groups) and the nucleotide differences between them, a substitution rate of 3.5×10^{-8} subs/site/year was estimated (Sakaoka et al. 1994). This rate is two-to-three orders of magnitude lower than any of the substitution rates we estimate here, including that of HSV-1 and VARV (table 2). Hughes et al. (2009) suggested previously that the HSV-1 rate of Sakaoka et al. (1994) might be too low given the substitution rates estimated for VARV. However, a gene-specific mutation rate of approximately 2×10^{-8} mutations/site/replication has been experimentally derived for the thymidine kinase gene of HSV-1 (the same gene analyzed here), an estimate congruent with both the substitution rate estimated by Sakaoka et al. (1994), and for dsDNA viruses generally (Drake 1991, Lu et al. 2002, Drake and Hwang 2005).

If the mutation rate of HSV-1 is 2×10^{-8} mutations/site/replication, we must then ask if it is possible for this rate to be translated into the substitution rate of 10^{-5} subs/site/year we estimated? The most likely mechanism through which such an inflated substitution rate could occur is through strong and continuous positive selection. To test for the presence of positive selection within our HSV-1 thymidine kinase gene alignment, the overall ratio of nonsynonymous to synonymous substitutions per site (d_N/d_S) was estimated using the program HyPhy (Kosakovsky Pond et al. 2005). The d_N/d_S ratio for this data set was 0.700, consistent with either localized positive selection on some codons or perhaps a substantial relaxation of purifying selection on this gene, as has been previously suggested (Drake and Hwang 2005). Hence, although positive selection may be contributing to the high substitution rates estimated for this gene, it seems highly unlikely that such adaptive evolution could result in a substitution rate that is three orders of magnitude higher than expected. We are therefore unable to conclusively reconcile the high substitution rates estimated here for HSV-1 (and potentially the related VZ virus), with what is known about the background mutation rate and biology of herpesviruses. At the very least, our HSV-1 analysis strongly suggests that future research should focus on acquiring sufficient herpesvirus data to allow for a rigorous estimate of the substitution and replication rates of these viruses, independent from any assumption of codivergence.

High Rates of Evolution in dsDNA viruses?

Clearly, the high rates of molecular evolution we have measured here for all seven dsDNA viruses are not easily explained by the understood biology of these (and other) dsDNA viruses. However, we cannot simply attribute the remarkably similar rate estimates across the seven viruses to an error in the method of estimation, as we were unable to recover comparably high estimates from the analysis of our synthetic data sets. While we have little confidence in the rates estimated for some of our data sets (HAdV-C, VZ, HPV-16), others are suggestive of high substitution rates in dsDNA viruses (VARV, HAdV-B, and perhaps HSV-1 and BK) and clearly merit further investigation. It is possible to identify two general mechanisms that may be driving the substitution rate of these viruses above the background mutation rate (~2×10^{-8} to 7×10^{-7}

mutations/site/replication, Drake et al. 1998). First, as the rate of viral replication directly impacts the long-term substitution rate (this parameter is analogous to generation time), very high replication rates have the potential to inflate the substitution rates of dsDNA viruses (McLysaght et al. 2003, Gubser et al. 2004, Esposito et al. 2006, Hughes et al. 2009). The impact of high replication rates may be particularly important for those viruses that are highly transmissible and/or result in acute infections in humans (VARV, HAdV-B), but are unlikely to have a similar effect in those viruses causing asymptomatic, latent and/or chronic infections (BK, HSV-1) in the host. Unfortunately, the replication rates of these viruses during natural infections in the host are unknown. Second, as discussed in the context of HSV-1, strong (diversifying) positive selection could also act to increase the substitution rate above the (neutral) mutation rate. Positive selection is most likely to be observed in viruses that cause a strong immune response in the host, or are the target of intense vaccination or intervention campaigns. However, as with replication rates, relatively little is known about the extent and nature of adaptive evolution in dsDNA viruses, although to date positive selection has not been identified in the protein coding genes of VARV (Hughes et al. 2009).

More generally, understanding the mechanisms of dsDNA virus evolution is central to the accurate assessment of these infectious agents as potentially emerging diseases of both humans and animals, particularly as the focus thus far has been primarily directed toward RNA viruses. Investigating the rates and processes of dsDNA virus evolution independent of the hypothesis of codivergence therefore constitutes an important avenue for future research.

Acknowledgements

CF received funding from the Natural Sciences and Engineering Research Council of Canada. MAS is support by United States Public Health Service grant GM086887. We thank the National Evolutionary Synthesis Center for hosting the working group “Software for Bayesian Evolutionary Analysis by Sampling Trees”, from which this project profited.

Literature Cited

Babkin IV, Shchelkunov SN. 2008. Molecular evolution of poxviruses. Russ J Genet. 44:895-908.

Beer BE, Bailes E, Goeken R, et al. (11 co-authors). 1999. Simian immunodeficiency virus (SIV) from sun-tailed monkeys (*Cercopithecus solatus*): Evidence for host-dependent evolution of SIV within the *C. lhoesti* superspecies. J Virol. 73: 7734-7744.

Benkö M, Harrach B. 2003. Molecular evolution of adenoviruses. Curr Top Microbiol Immunol. 272:3-35.

Bernard H-U. 1994. Coevolution of papillomaviruses with human populations. Trends Microbiol. 2:140-143.

Bernard H-U, Calleja-Macias IE, Dunn ST. 2006. Genome variation of human papillomavirus types: phylogenetic and medical implications. Int J Cancer. 11:1071-1076.

Bryant JE, Holmes EC, Barrett AD. 2007. Out of Africa: a molecular perspective on the introduction of yellow fever virus into the Americas. PLoS Path. 3:e75.

Charrel RM, De Micco P, de Lamballerie X (1999) Phylogenetic analysis of GB viruses A and C: evidence for codivergence between virus isolates and their primate hosts. J Gen Virol 80: 2329-2335.

Chen Y, Sharp PM, Fowkes M, Kocher O, Joseph JT, Koralnik IJ. 2004. Analysis of 15 novel full-length BK virus sequences from three individuals: evidence of a high intra-strain genetic diversity. J Gen Virol. 85:2651-2663.

Chen Z, DeSalle R, Schiffman M, Herrero R, Burk RD. 2009. Evolutionary dynamics of variant genomes of human papillomavirus types 18, 45 and 97. J Virol. 83:1443-1455.

de Oliveira T, Pybus OG, Rambaut A, et al. (14 co-authors). 2006. Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. Nature. 444:836-837.

Drake JW. 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proc Natl Acad Sci USA. 88:7160-7164.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics. 148:1667-1686.

Drake JW and Hwang CBC. 2005. On the mutation rate of herpes simplex virus type 1. Genetics. 170:969-970.

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

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.

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. 7:214-221.

Duffy S, Holmes EC. 2009. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. J Gen Virol. 90:1539-1547.

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

Emerson BC. 2007. Alarm bells for the molecular clock? No support for Ho et al.'s model of time-dependent molecular rate estimates. Syst Biol. 56:337-345.

Esposito JJ, Sammons SA, Frace AM, et al. (14 co-authors). 2006. Genome sequence diversity and clues to the evolution of variola (smallpox) virus. Science. 313:807-812.

Firth C, Charleston MA, Duffy S, Shapiro B, Holmes EC. 2009. Insights into the evolutionary history of an emerging livestock pathogen: Porcine Circovirus 2. J Virol. 83:12813-12821.

Gubser C, Hué S, Kellam P, Smith GL. 2004. Poxvirus genomes: a phylogenetic analysis. J Gen Virol. 85:105-117.

Harkins GW, Delport W, Duffy S, et al. (14 co-authors). 2009. Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. Virol J. 16:104-117.

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.

Ho SYW and Larson G. 2006. Molecular clocks: when times are a-changin'. Trends Genet. 22:79-83.

Ho SYW, Shapiro B, Phillips MJ, Cooper A, Drummond AJ. 2007a. Evidence for time dependency of molecular rate estimates. Syst Biol. 56:515-522.

Ho SYW, Kolokotronis S-O, Allaby RG. 2007b. Elevated substitution rates estimated from ancient DNA sequences. Biol Lett. 3:702-705.

Ho SYW, Saarma U, Barnett R, Haile J, Shapiro B. 2008. The effect of inappropriate calibration: three case studies in molecular ecology. PLoS One. 3:e1615.

Holmes EC. 2004. The phylogeography of human viruses. Mol Ecol. 13:745-756.

Holmes EC, Drummond AJ. 2007. The evolutionary genetics of viral emergence. Curr Top Microbiol Immunol. 315:51-66.

Hughes AL, Friedman R. 2000. Evolutionary diversification of protein-coding genes of hantaviruses. Mol Biol Evol. 17:1558-1568.

Hughes AL, Irausquin S, Friedman. 2009. The evolutionary biology of poxviruses. Infect Genet Evol. doi:10.1016/j.meegid.2009.10.001.

Jackson AP. 2005. The effect of paralogous lineages on the application of reconciliation analysis by cophylogeny mapping. Syst Biol. 54:127-145.

Jenkins GM, Rambaut A, Pybus OG, Holmes EC. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetics analysis. J Mol Evol. 54:152-161.

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

Kosakovsky Pond SL, Frost SDW, Muse SD. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 21:676-679.

Krumbholz A, Bininda-Emonds OR, Wutzler P, Zell R. 2008. Evolution of four BK virus subtypes. Infect Genet Evol. 8:632-643.

Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots. PLoS Comp Biol. 5:e1000520.

Lewis-Rogers N, Crandall KA. 2009. Evolution of picornaviridae: an examination of phylogenetic relationships and cophylogeny. Mol Phylogenet Evol. doi:10.1016/j.ympev.2009.10.015.

Li Y, Carroll DS, Gardner SN, Walsh MC, Vitalis EA, Damon IK. 2007. On the origin of smallpox: correlating variola phylogenetics with historical smallpox records. Proc Natl Acad Sci USA. 104:15787-15792.

Lu Q, Hwang YT, Hwang CBC. 2002. Mutation spectra of herpes simplex virus type 1 thymidine kinase mutants. J Virol. 76:5822-5828.

Madisch I, Harste G, Pommer H, Heim A. 2005. Phylogenetic analysis of the main neutralization and hemagglutination determinants of all human adenovirus prototypes as a basis for molecular classification and taxonomy. J Virol. 79:15265-15276.

Martin DP, Williamson C, Posada D. 2005. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics. 21:260-262.

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 EAR. 1995. Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. J Mol Biol. 247: 443-458.

McGeoch DJ, Dolan A, Ralph AC. 2000. Toward a comprehensive phylogeny for mammalian and avian herpesviruses. J Virol. 74: 10401-10406.

McLysaght A, Baldi PF, Gaut BS. 2003. Extensive gene gain associated with adaptive evolution of poxviruses. Proc Natl Acad Sci USA. 100:15655-15660.

Muir WB, Nichols R, Breuer J. 2002. Phylogenetic analysis of varicella-zoster virus:evidence of intercontinental spread of genotypes and recombination. J Virol. 76:1971-1979.

Nakao H, Okomoto H, Fukuda M, Tsuda F, Mitsui T, Masuko K, Iizuka H, Miyakawa Y, Mayumi M. 1997. Mutation rate of GB Virus C/hepatitis G virus over the entire genome and in subgenomic regions. Virology. 233:43-50.

Navascues M, Emerson BC. 2009. Elevated substitution rate estimates from ancient DNA:model violation and bias of Bayesian methods. Mol Ecol. 18:4390-4397.

Nishimoto Y, Takasaka T, Hasegawa M, Zheng H-Y, Chen Q, Sugimoto C, Kitamura T, Yogo Y. 2006. Evolution of BK virus based on complete genome data. J Mol Evol. 63:341-352.

Nemirov K, Vaheri A, Plyusnin A. 2004. Hantaviruses: co-evolution with natural hosts. Recent Res Devel Virol. 6:201-228.

Ong C-K, Chan S-Y, Camp MS, et al. (11 co-authors). 1993. Evolution of human papillomavirus type 18:an ancient phylogenetic root in Africa and intratype diversity reflect coevolution with human ethnic groups. J Virol. 67:6424-6431.

Peiris JS, Guan Y, Yuen KY. 2004. Severe acute respiratory syndrome. Nature Medicine. 10:S88-97.

Penny D. 2005. Relativity for molecular clocks. Nature. 336:183-184.

Pérez-Losada M, Christensen RG, McClellan DA, Adams BJ, VIscidi RP, Demma JC, Crandall KA. 2006. Comparing phylogenetic codivergence between polyomaviruses and their hosts. J Virol. 80:5663-5669.

Pérez-Losada M, Crandall KA, Zenilman J, Viscidi RP. 2007. Temporal trends in gonococcal population genetics in a high prevalence urban community. Infect Genet Evol. 7:271-278.

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics. 9:817-818.

Pybus OG, Rambaut A, Harvey PH. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics. 155:1429-1437.

Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. 2008. The genomic and epidemiological dynamics of human influenza A virus. Nature. 453: 615-619.

Ramsden C, Holmes EC, Charleston MA. 2009. Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. Mol Biol Evol. 26:143-153.

Rector A, Lemey P, Tachezy R, et al. (15 co-authors). 2007. Ancient papillomavirus –host co-speciation in Felidae. Genome Biol. 8:R57.

Romano CM, Zanotto PM de A, Holmes EC. 2008. Bayesian coalescent analysis reveals a high rate of molecular evolution in GC virus C. J Mol Evol. 66:292-297.

Roy S, Vandenberghe LH, Kryazhimskiy S, et al. (12 co-authors). 2009. Isolation and characterization of adenoviruses persistently shed from the gastrointestinal tract of non-human primates. PLoS Path. 5:e1000503.

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. 75:513-527.

Shackelton LA, Parrish CR, Truyen U, Holmes EC. 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc Natl Acad Sci USA. 102:379-384.

Shackelton LA, Rambaut A, Pybus OG, Holmes EC. 2006. JC virus evolution and its association with human populations. J Virol. 80:9928-9933.

Shchelkunov SN. 2009. How long ago did smallpox virus emerge? Arch Virol. 154:1865-1871.

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 USA. 106:11709-11712.

Strimmer K, Pybus OG. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. Mol Biol Evol. 18:2298-2305.

Suchard MA, Weiss RE, Sinsheimer JS. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol Biol Evol. 16:1001-1013.

Sugimoto C, Hasegawa M, Kato A, Zheng H-Y, Ebihara H, Taguchi F, Kitamura T, Yogo Y. 2002. Evolution of human polyomavirus JC: implications for the population history of humans. J Mol Evol. 54:285-297.

Villarreal LP, Defilippis VR, Gottlieb KA. 2000. Acute and persistent viral life strategies and their relationship to emerging diseases. Virology. 272:1-6.

Wagenaar TR, Chow VTK, Buranathai C, Thawatsupha P, Grose C. 2003. The out of Africa model of varicella-zoster evolution: single nucleotide polymorphisms and private alleles distinguish Asian clades from European/North American clades. Vaccine. 21:1072-1081.

Yasunaga T, Miyata T. 1982. Evolutionary changes of nucleotide sequences of papova viruses BKV and SV40: they are possibly hybrids. J Mol Evol. 19:72-79.

Zheng H-Y, Nishimoto Y, Chen Q, et al. (11 co-authors). 2007. Relationships between BK virus lineages and human populations. Microbes Infect. 9:204-213.

Table 1. The gene, number of sequences, length (number of base pairs) and sampling interval for each of the dsDNA virus used in this analysis.

Virus	Gene	No. of taxa	Length (bp)	Sampling Range
HPV-16	L1 protein	34	1510	2005 - 2008
HSV-1	Thymidine kinase	84	1135	1981 - 2008
HAdV-B	Hexon	115	2835	1983 - 2007
HAdV-C	Hexon	71	1419	1986 - 2007
VZ	Genome	20	126563	1970 - 2007
VARV	Genome	47	185600	1946 - 1977
BK	Genome (coding only)	212	4838	1978 - 2007

Table 2. Mean and 95% highest probability density (HPD) of the Bayesian posterior estimates of substitution rate (subs/site/year) and time to most recent common ancestor (TMRCA) (ybp) for seven dsDNA viruses: Human Papillomavirus Type-16, Herpes Simplex Virus-1, Variola Virus (VZ), BK virus (BK), Variola virus (VARV), Human Adenovirus Subtype B (HAdV-B) and Human Adenovirus Subtype C (HAdV-C). Estimates were made under a variety of molecular clock models (strict, relaxed with a lognormal distribution of rates, and relaxed with an exponential distribution of rates), as well as under two demographic models (constant population size and Bayesian Skyline (BS)). The best model was chosen as the one with the highest marginal likelihood (Suchard et al. 2001), and is indicated with an asterisk (*).

Virus	Clock Model	Demographic	Marginal	Mean Rate	Substitution Rate	TMRCA	TMRCA
		Model	Likelihood	(subs/site/year)	95% HPD	(ybp)	95% HPD
HPV-16							
	Strict	Constant	-2636	4.94x10^{-4}	0.7490 – 9.812x10^{-4}	22	7 – 46
		BS	-2638	4.27x10^{-4}	0.0953 – 9.194x10^{-4}	30	6 – 67
	UCLN	Constant	-2619	4.85x10^{-3}	0.0001 – 14.91x10^{-3}	46	3 – 27
		BS*	-2618	3.94x10^{-3}	0.0175 – 14.23x10^{-3}	6	3 - 15
	UCED	Constant	-2624	8.78x10^{-4}	0.0494 – 2.164x10^{-3}	18	4 - 44
		BS	-2624	7.29x10^{-4}	0.0009 – 1.856x10^{-3}	23	3 – 53
HSV-1							
	Strict	Constant	-2522	8.65x10^{-5}	0.4995 – 1.264x10^{-4}	140	83 – 212
		BS	-2512	8.79x10^{-5}	0.5727 – 1.208x10^{-4}	133	84 – 191
	UCLN	Constant	-2521	8.69x10^{-5}	0.5027 – 1.274x10^{-4}	144	81 - 226
		BS	-2511	8.75x10^{-5}	0.5457 – 1.215x10^{-4}	137	79 – 202
	UCED	Constant	-2516	8.96x10^{-5}	0.4386 – 1.398x10^{-4}	183	66 – 356
		BS*	-2508	8.21x10^{-5}	0.4387 – 1.269x10^{-4}	194	62 – 387
VZ							
	Strict	Constant	-182462	3.80x10^{-6}	2.192 – 5.531x10^{-6}	322	186 – 452
		BS	-182463	3.77x10^{-6}	2.010 – 5.480x10^{-6}	303	183 – 466
	UCLN	Constant	-182423	4.75x10^{-6}	1.764 – 8.435x10^{-6}	286	92 - 546

		BS	-182423	5.48x10^{-6}	1.859 – 9.487x10^{-6}	235	74 – 431
	UCED	Constant	-182421	4.96x10^{-6}	0.0697 – 2.050x10^{-5}	516	70 – 1031
		BS*	-182421	6.26x10^{-6}	0.0694 – 1.235x10^{-5}	309	51 – 741
BK							
	Strict	Constant	-14623	3.29x10^{-5}	2.307 – 4.311x10^{-5}	1033	734 - 1386
		BS	-14619	2.15x10^{-5}	0.8400 – 3.466x10^{-5}	1740	791 – 3180
	UCLN	Constant	-14573	3.89x10^{-5}	2.515 – 5.371x10^{-5}	933	436 - 1511
		BS	-14569	3.01x10^{-5}	1.344 – 4.681x10^{-5}	1227	442 - 2200
	UCED	Constant	-14550	4.77x10^{-5}	2.877 – 6.700x10^{-5}	770	303 - 1394
		BS*	-14548	4.34x10^{-5}	2.416 – 6.411x10^{-5}	824	286 - 1550
VARV							
	Strict	Constant	-261780	9.03x10^{-6}	7.847 – 10.20x10^{-6}	192	167 - 218
		BS	-261779	8.98x10^{-6}	7.772 – 10.16x10^{-6}	193	168 - 220
	UCLN	Constant	-261761	8.27x10^{-6}	5.889 – 10.53x10^{-6}	217	136 – 313
		BS	-261760	8.21x10^{-6}	5.905 – 10.48x10^{-6}	215	136 - 309
	UCED	Constant	-261743	9.10x10^{-6}	5.157 – 13.09x10^{-6}	206	85 - 366
		BS*	-261742	9.32x10^{-6}	4.977 – 13.84x10^{-6}	197	76 – 372
HAdV-B							
	Strict	Constant	-4631	6.91x10^{-5}	0.3949 – 1.019x10^{-4}	47	31 - 68
		BS	-4627	6.87x10^{-5}	3.877 – 9.981x10^{-5}	43	27 - 62
	UCLN	Constant	-4598	6.86x10^{-5}	0.3479 – 1.043x10^{-4}	61	28 - 111
		BS*	-4598	7.20x10^{-5}	0.3661 – 1.113x10^{-4}	50	24 - 98
	UCED	Constant	-4608	7.19x10^{-5}	0.383 – 1.066x10^{-4}	53	29 – 91
		BS	-4607	7.47x10^{-5}	0.4087 – 1.125x10^{-4}	44	25 – 75
HAdV-C							
	Strict	Constant	-2294	3.34x10^{-5}	0.5132 – 6.299x10^{-5}	354	71 – 591
		BS	-2290	2.44x10^{-5}	0.1006 – 5.146x10^{-5}	405	62 – 937
	UCLN	Constant	-2285	4.31x10^{-5}	0.9117 – 8.236x10^{-5}	190	32 – 441
		BS	-2285	4.23x10^{-5}	0.3472 – 8.545x10^{-5}	189	22 - 503
	UCED	Constant	-2287	4.01x10^{-5}	1.009 – 7.528x10^{-5}	233	45 - 444
		BS*	-2284	3.46x10^{-5}	0.0139 – 6.798x10^{-5}	297	37 - 628

Figure Legends

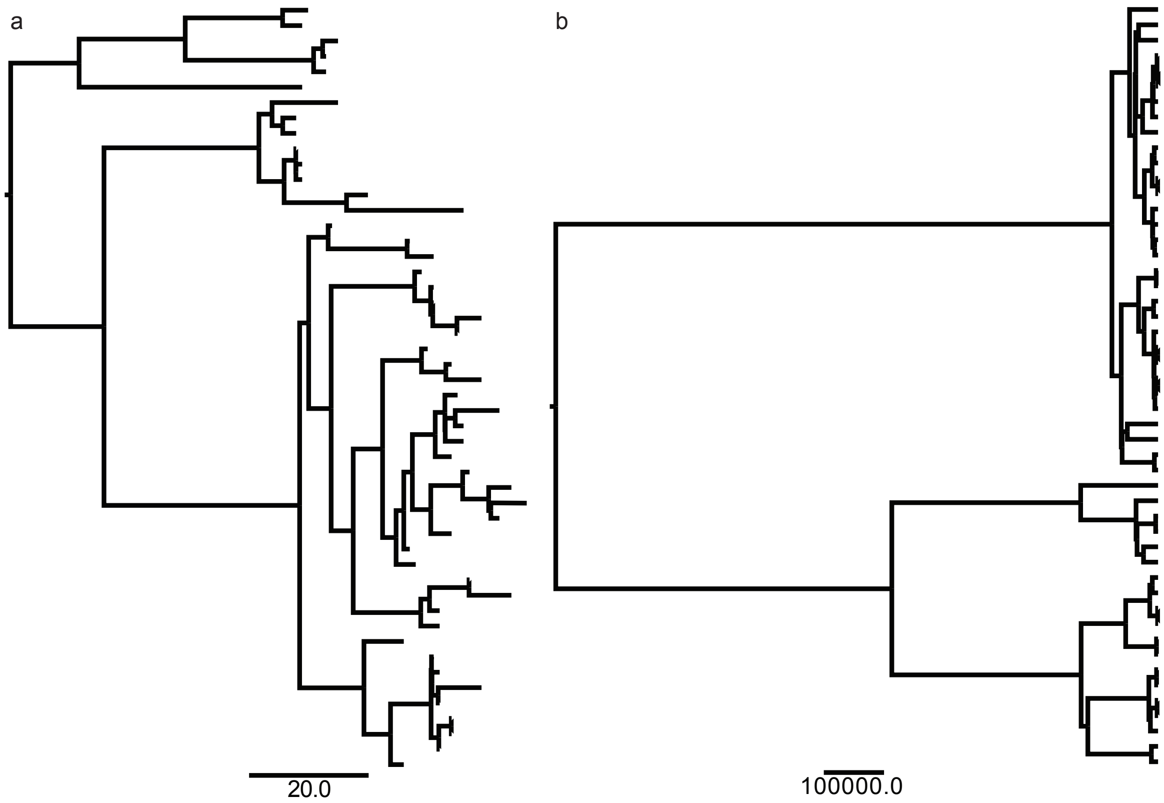
Figure 1. Tree diagrams with identical taxa numbers sampled over identical time intervals. When the sampling interval is similar to the time frame over which sequence evolution occurs (10^{-4} subs/site/year), it is possible to assess the long-term rate of evolution with high precision (a). When the sampling interval is small relative to the time frame of sequence evolution (10^{-8} subs/site/year), it may become difficult to accurately estimate substitution rates (b). The scale bars indicate the branch lengths in number of years.

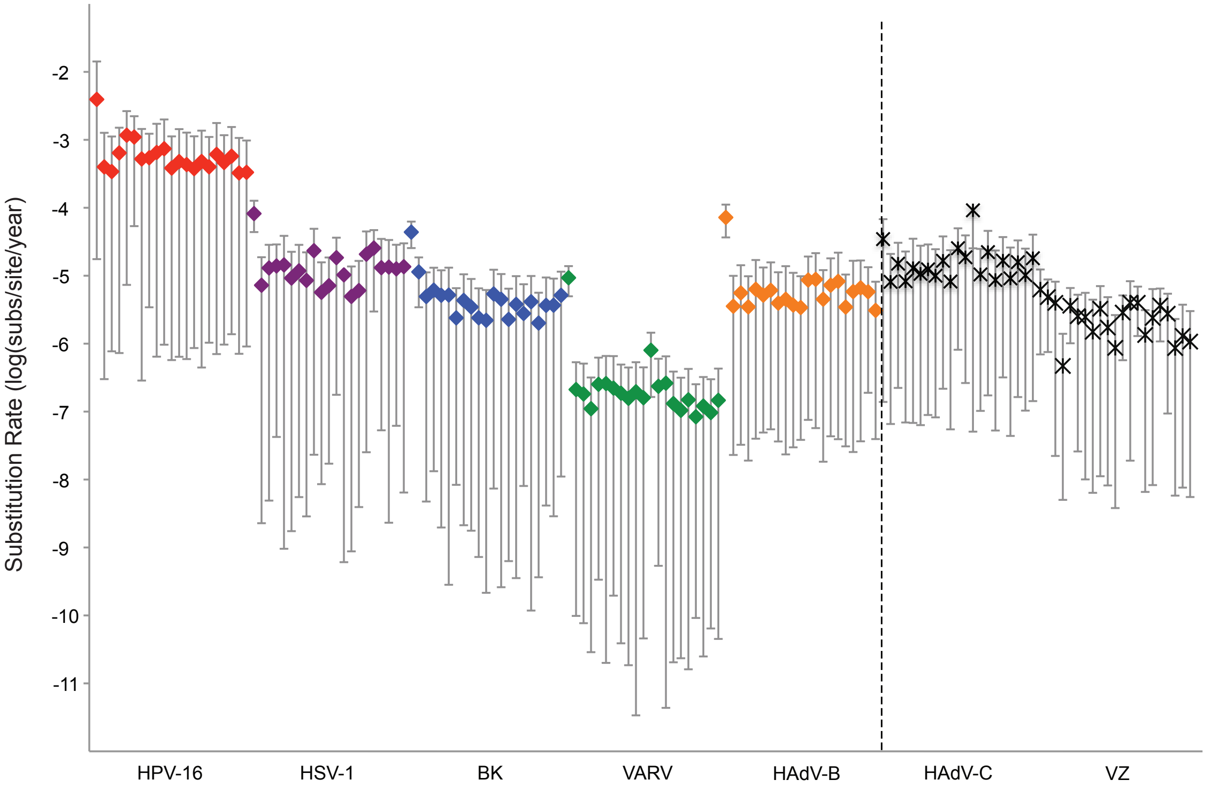
Figure 2. Posterior mean and 95% HPDs of the substitution rates estimated from the actual data sets (far left value for each virus) and the 20 tip-date randomizations for the dsDNA viruses Human Papillomavirus Type-16 (HPV-16), Herpes Simplex Virus-1 (HSV-1), BK virus (BK), Variola virus (VARV), Human Adenovirus Subtype-B (HAdV-B), Human Adenovirus Subtype-C (HAdV-C) and Varicella Zoster virus (VZ). The mean rates estimated for the HAdV-C and VZ data sets were not significantly different from those estimated from the randomized data sets.

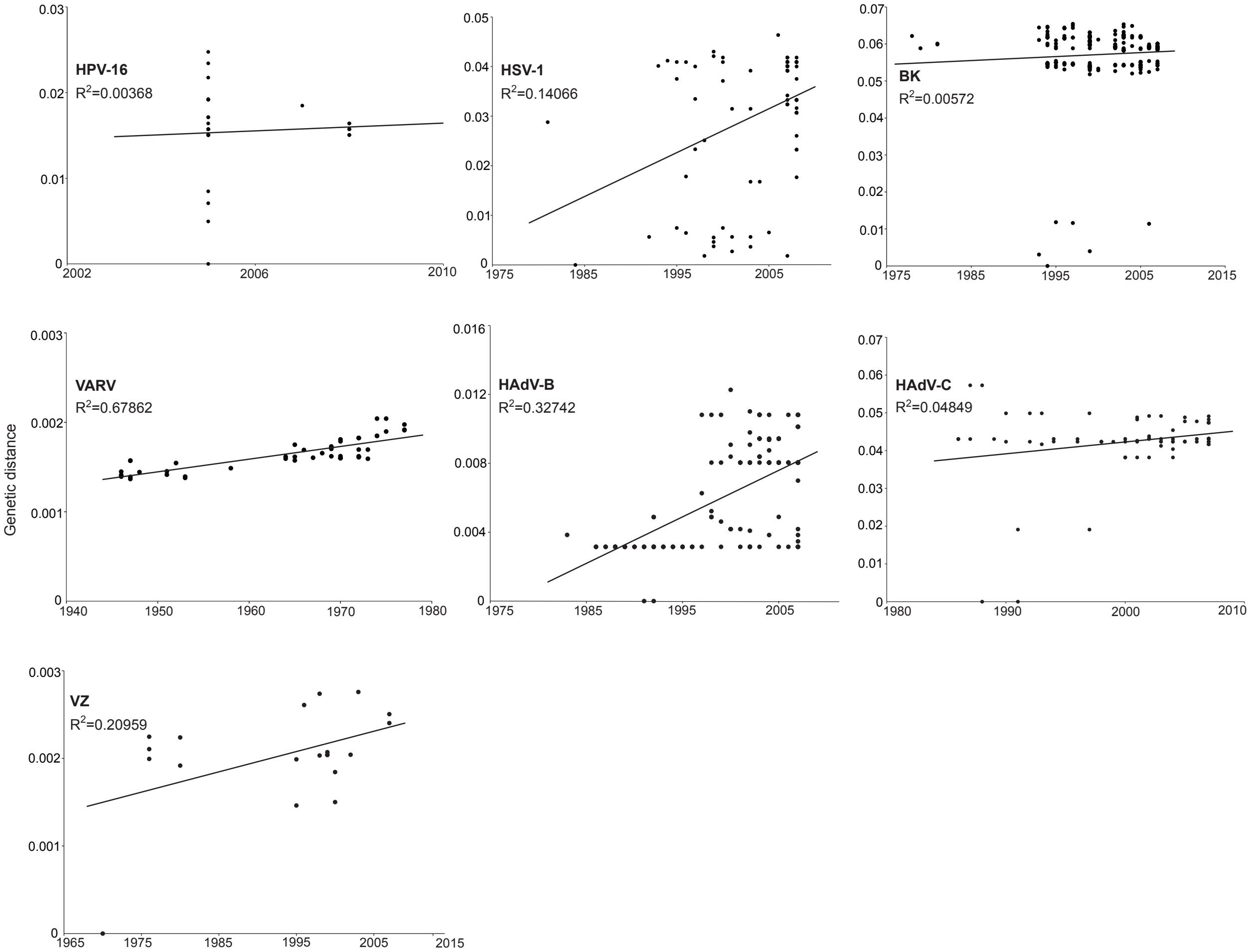
Figure 3. Genetic distance versus sampling year for the dsDNA viruses (clockwise from top left): Human Papillomavirus Type-16 (HPV-16), Herpes Simplex Virus-1 (HSV-1), BK Virus (BK), Variola Virus (VARV), Human Adenovirus Subtype B (HAdV-B), Human Adenovirus Subtype C (HAdV-C), and Varicella Zoster Virus (VZ). The regression coefficient (R^2) estimates the fit of the data to a strict molecular clock by testing the degree of influence sampling time has over the amount of pairwise diversity in the data. This analysis supports the presence of temporal structure in the data for VARV and HAdV-B, while suggesting the presence of temporal structure for HSV-1 and VZ. No evidence for temporal structure within the sampled period was found for the HPV-16, BK and HAdV-C data sets using this method.

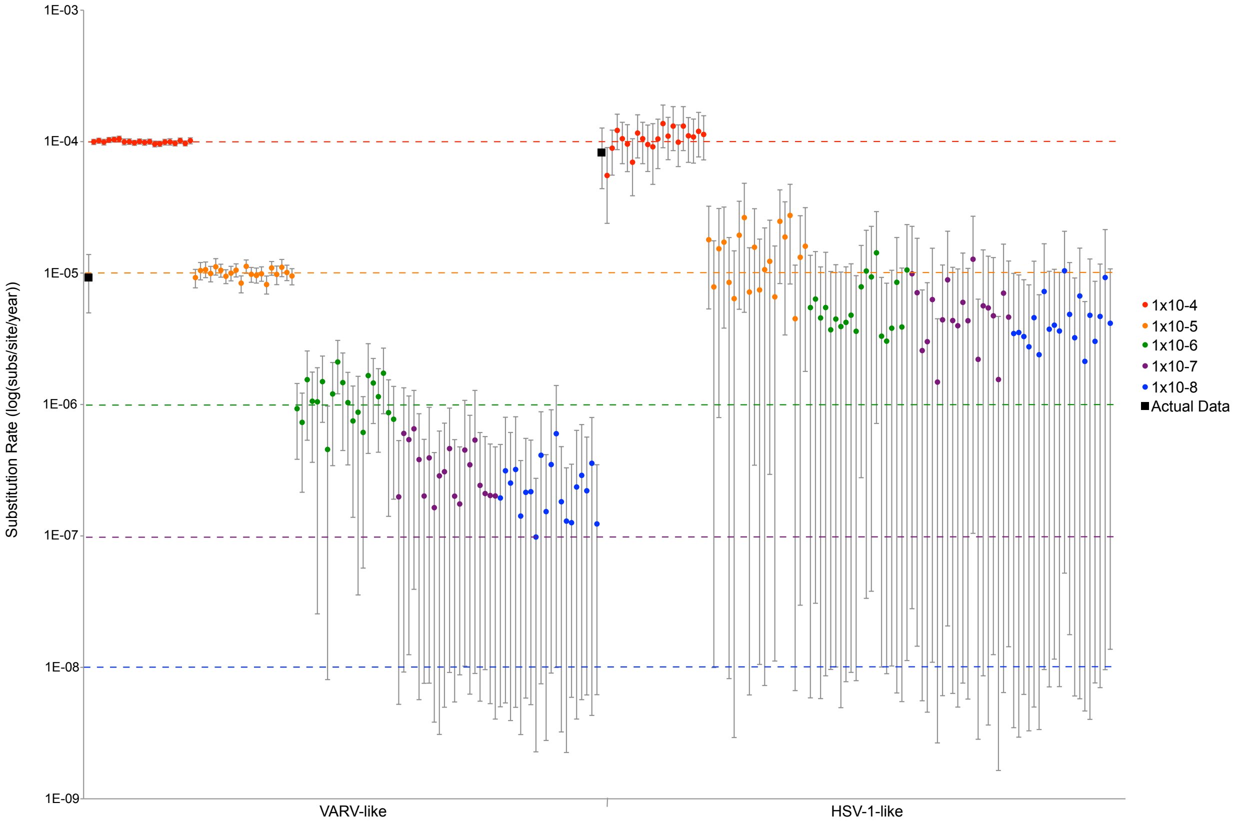
Figure 4. Substitution rates (posterior mean and 95% HPD) estimated from synthetic data sets based on the VARV and HSV-1 data sets from the first part of this study. Twenty replicates of both the VARV- and HSV-1-like data sets were generated under strict molecular clocks evolving at each of 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7} and 10^{-8} subs/site/year. The dashed lines show the true mean evolutionary rate for each group of simulations of the corresponding color. The mean and 95% HPDs of the substitution rates for VARV and HSV-1 under the best evolutionary model are shown for comparison.

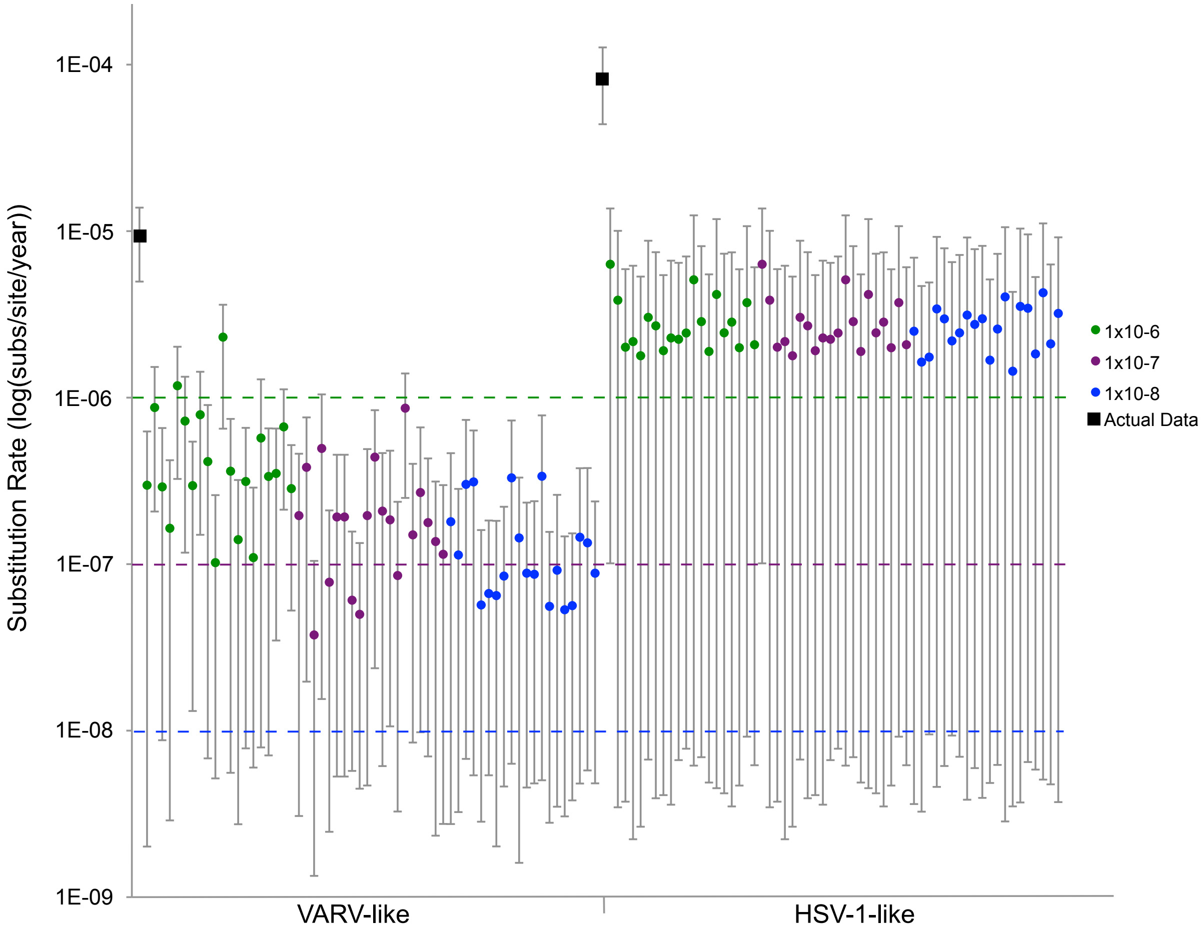
Figure 5. Substitution rates (posterior mean and 95% HPD) estimated from simulated data sets based on the VARV and HSV-1 data sets from the first part of this study. Twenty replicates of both the VARV- and HSV-1-like data sets were generated under relaxed molecular clocks evolving at each of 10^{-6}, 10^{-7} and 10^{-8} subs/site/year. The dashed lines show the true mean evolutionary rate for each group of simulations of the corresponding color. The mean and 95% HPDs of the substitution rates for VARV and HSV-1 under the best evolutionary model are shown for comparison.


a
b
20.0
100000.0


Substitution Rate (log(subs/site/year))
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
HPV-16
HSV-1
BK
VARV
HAdV-B
HAdV-C
VZ


Genetic distance
HPV-16
R^2=0.00368
0.03
0.02
0.01
0
2002
2006
2010
HSV-1
R^2=0.14066
0.05
0.04
0.03
0.02
0.01
0
1975
1985
1995
2005
BK
R^2=0.00572
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1975
1985
1995
2005
2015
VARV
R^2=0.67862
0.003
0.002
0.001
0
1940
1950
1960
1970
1980
HAdV-B
R^2=0.32742
0.016
0.012
0.008
0.004
0
1975
1985
1995
2005
HAdV-C
R^2=0.04849
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1980
1990
2000
2010
VZ
R^2=0.20959
0.003
0.002
0.001
0
1965
1975
1985
1995
2005
2015


1E-03
1E-04
1E-05
1E-06
1E-07
1E-08
1E-09
Substitution Rate (log(subs/site/year))
VARV-like
HSV-1-like
● 1x10-4
● 1x10-5
● 1x10-6
● 1x10-7
● 1x10-8
■ Actual Data


1E-04
1E-05
1E-06
1E-07
1E-08
1E-09
Substitution Rate (log(subs/site/year))
VARV-like
HSV-1-like
1x10-6
1x10-7
1x10-8
Actual Data