LA-UR- 09-06062

Approved for public release; distribution is unlimited.

Title:	The evolutionary rate dynamically tracks changes in HIV-1 epidemics
Author(s):	Irina Maljkovic Berry, Z#: 203158, T-6/T Division Gayathri Athreya, Z#: 218561, T-6/T Division Marcus Daniels, Z#: 211500, T-6/T Division William J. Bruno, Z#: 107647, T-6/T Division Bette Korber, Z#: 108817, T-6/T Division Carla Kuiken, Z#: 111147, T-6/T Division Ruy M. Ribeiro, Z#: 171295, T-6/T Division
Intended for:	Journal: Epidemics

Los Alamos
NATIONAL LABORATORY
EST. 1943

Form 836 (7/06)

Elsevier Editorial System(tm) for Epidemics
Manuscript Draft

Manuscript Number:

Title: The evolutionary rate dynamically tracks changes in HIV-1 epidemics

Article Type: Original Research

Keywords: Viral evolution
Molecular epidemiology
Phylogeny
TreeRate

Corresponding Author: Dr Thomas Leitner,

Corresponding Author's Institution: Los Alamos National Laboratory

First Author: Irina Maljkovic Berry

Order of Authors: Irina Maljkovic Berry; Gayathri Athreya; Moulik Kothari; Marcus Daniels; William J Bruno; Bette Korber; Carla Kuiken; Ruy M Ribeiro; Thomas Leitner

Abstract: Large sequence datasets provide an opportunity to investigate the dynamics of pathogen epidemics. Thus, a fast method to estimate the evolutionary rate from large and numerous phylogenetic trees becomes necessary. Based on minimizing tip height variances, we optimize the root in a given phylogenetic tree, to estimate the most homogenous evolutionary rate between samples from at least two different time points. Simulations showed that the method had no bias in the estimation of evolutionary rates, and that it was robust to tree rooting and topological errors. We show that the evolutionary rates of HIV-1 subtype B and C epidemics have changed over time, with the rate of evolution inversely correlated to the rate of virus spread. For subtype B the evolutionary rate slowed down and tracked the start of the HAART era in 1996. Subtype C in Ethiopia showed an increase in the evolutionary rate when the prevalence increase markedly slowed down in 1995.

Thus, we show that the evolutionary rate of HIV-1 on the population level dynamically tracks epidemic events.

Suggested Reviewers: Ron Swanstrom
risunc@med.unc.edu

Angela McLean
angela.mclean@zoo.ox.ac.uk

Opposed Reviewers: Alexei Drummond
conflict of interest

Andrew Rambaut
conflict of interest

Marc Suchard
conflict of interest

Philippe Lemey
conflict of interest

Cover Letter

Dear Sir,

Please find enclosed manuscript titled "A simple method for optimizing the root and evolutionary rate in phylogenetic trees with taxa collected at a minimum of two different time points" by Irina Maljkovic Berry, Gayathri Athreya, Moulik Kothari, Marcus Daniels, Bette Korber, Carla Kuiken, and Thomas Leitner, that we would like to submit for possible publication in Epidemics.

This paper describes a fast and accurate method to root and estimate evolutionary rates in a given phylogenetic tree. While there are several methods that estimate rates and that can root trees, no method combines them into a fast and simple strategy. Thus, our method can take a tree calculated with any tree building method and optimize the root and rate in it. It is especially useful for large analyses, with either many taxa or many datasets. We have evaluated our method on simulated data that aimed at investigating the performance under many limiting situations, and found it to perform very good under situations that are typical in biological systems.

We have used this method to analyze the HIV-1 subtype B and C epidemics. We show that our method is able to track dynamic changes in these epidemics, and thus that it may be applicable to infer and predict changes in epidemics involving pathogens that evolve during their spread. We believe these results and this method are of great interest to your readers.

The paper has not been sent to any other journal, and there are no other papers that currently are under consideration and relate to this paper from us. None of the authors have any conflicts of interest to delclare.

Sincerely,
Thomas Leitner, PhD

The evolutionary rate dynamically tracks changes in HIV-1 epidemics: application of a simple method for optimizing the root in phylogenetic trees with longitudinal data

Irina Maljkovic Berry[a, b, c+*], Gayathri Athreya [a+], Moulik Kothari[a], Marcus Daniels[a], William J. Bruno[a], Bette Korber[a], Carla Kuiken[a], Ruy M. Ribeiro[a], Thomas Leitner[a*]

[a] Theoretical Biology & Biophysics, MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.

[b] Center for Nonlinear Studies (CNLS), Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.

[c] Department of Virology, Swedish Institute for Infectious Disease Control, SE-171 82 Solna, & Department of Microbiology, Tumor and Cell Biology, Karolinska Institute, SE-171 77 Stockholm, Sweden

[+] These authors contributed equally to this study

[*] Author's full last name is Maljkovic Berry (double name)

Corresponding author: Thomas Leitner, tkl@lanl.gov Phone: +1-505-667-3898, Fax: +1-505-665-3493

ABSTRACT

Large sequence datasets provide an opportunity to investigate the dynamics of pathogen epidemics. Thus, a fast method to estimate the evolutionary rate from large and numerous phylogenetic trees becomes necessary. Based on minimizing tip height variances, we optimize the root in a given phylogenetic tree, to estimate the most homogenous evolutionary rate between samples from at least two different time points. Simulations showed that the method had no bias in the estimation of evolutionary rates, and that it was robust to tree rooting and topological errors. We show that the evolutionary rates of HIV-1 subtype B and C epidemics have changed over time, with the rate of evolution inversely correlated to the rate of virus spread. For subtype B the evolutionary rate slowed down and tracked the start of the HAART era in 1996. Subtype C in Ethiopia showed an increase in the evolutionary rate when the prevalence increase markedly slowed down in 1995. Thus, we show that the evolutionary rate of HIV-1 on the population level dynamically tracks epidemic events.

Keywords: Viral evolution, Molecular epidemiology, Phylogeny, TreeRate

INTRODUCTION

The rate of evolution is a fundamental quantity in the field of molecular biology and evolution, and has often been measured as the rate of nucleotide substitutions. Estimating the rate of substitutions is especially effective when there are known dates not only at the tips of a phylogenetic tree, but also deeper into the tree. This situation exists when there are either fossil data that can date historic events, or when the organism under study evolves fast enough to accumulate mutations for a researcher to sample it within reasonable time. The latter is the case among many viruses, where samples taken only a few years apart may display as much evolution as higher organisms do in millions of years (Leitner, 2002; Leitner and Albert, 1999). For example, HIV-1 evolution has been estimated at rates between $1x10^{-3}$ and $17x10^{-3}$ substitutions site^{-1} year^{-1} in *env* (Korber et al., 2000; Leitner et al., 1999; Maljkovic Berry et al., 2007; Salemi et al., 2001).

Various methods have been proposed to estimate the rate of substitutions over time, i.e., the molecular clock. Originally, the molecular clock was estimated as a constant accumulation of substitutions over time (Kimura, 1980; Zuckerkandl et al., 1965) but that simplifying assumption may not always be appropriate (Gillespie, 1984; Gillespie, 1988; Takahata, 1987) and more recently several Bayesian methods have been suggested on how to relax the strict molecular clock (Drummond et al., 2006; Huelsenbeck et al., 2000; Kishino et al., 2001; Sanderson, 2002; Thorne et al., 1998; Yang et al., 2006). Some other recent methods also allow for samples with different collection dates (Rodrigo et al., 2003), and yet other methods have investigated and incorporated uncertainties in the time stamps (Korber et al., 2000; Leitner et al., 1999; Yang et al., 2006). Furthermore, local molecular clocks that can

accommodate higher levels of rate heterogeneity than the Bayesian approaches have been developed (Aris-Brosou, 2007; Yoder et al., 2000). While the relaxed clocks in many cases appear to be more realistic and improve the rate estimates, they become more complex, requiring more assumptions to be made and more parameters to be estimated, and slow to run on computers. Also, a tree reconstructed under a fully unrestricted rate model, i.e., a tree with no clock assumption, may still have a better fit than a tree assuming a particular clock model. For these reasons, we have developed a fast and simple method to find the root that gives the most homogeneous rate in a given tree with samples from at least two different time points. The tree can be calculated by any method, and as long as the branch lengths are realistic measures of divergence, an average rate can be estimated for the time interval between the samples.

We apply this method to the epidemics of HIV-1 subtypes B and C, from Europe and North America, and Africa, respectively. We show that, for subtype B, the evolutionary rate is constant until 1997, after which a significant decrease in the rate is observed. Interestingly, this decrease coincides with the global onset of HAART in 1996. Furthermore, we did not observe a low evolutionary rate of the virus in the early epidemic, indicating that the period of exponential growth in the U.S.A. precedes most of the early documented sequences. Subtype C displayed large fluctuations. As in the subtype B epidemic, different countries in the subtype C epidemic had very different prevalence dynamics. Analyses of the Ethiopian subtype C sub-epidemic revealed an inverse correlation between virus spread and the evolutionary rate of HIV-1, where the evolutionary rate increased after 1995 when the rate of spread slowed down. Thus, we show that changes in HIV-1 epidemic can be revealed by

consecutively estimating the evolutionary rate.

METHODS

Root optimization

A standard Newick formatted tree is the input. The operational taxonomic units (OTUs) in the tree can be divided into two longitudinal samples, each with an average distance to the root $\overline{X}_i$ and separated by a time interval Δt. The distance between these samples is calculated as $\Delta\hat{d} = \overline{X}_2 - \overline{X}_1$ (Fig 1). It is also possible to use an additional discard group, where one can put sequences not to be considered in the $\Delta\hat{d}$ calculation. In that way, OTUs in one phylogenetic tree can be rearranged and reanalyzed in several different ways. This also allows for trees constructed with samples from more than two time points to be analyzed, *e.g.,* OTUs from a third (or many) time point(s) can be put in the discard group while $\Delta\hat{d}$ is calculated between time points one and two, then OTUs from time point one are put in the discard group and $\Delta\hat{d}$ is calculated between OTUs of time points two and three. Similarly, the method could be extended to optimize the tip height variances from all time points simultaneously (Eq.1). Thus, the method we propose measures the distance (amount of evolution) *between* OTUs in sample 1 and sample 2. It is primarily intended to estimate the evolutionary rate of a population sampled at (at least) two time points. For this calculation to be most reliable, sample 1 and 2 OTUs should preferably not be separated into two monophyletic groups but rather intermixed. This is because if the two samples were monophyletically divided then 1) biologically and epidemiologically, one could not be

certain that the two samples came from the same population or outbreak, and 2) mathematically, there would be no information on where to root the tree along the branch that separated the two samples because the variance would not change along it. Since $\Delta\hat{d}$ between the samples can differ depending on how the tree is rooted, $\Delta\hat{d}$ is calculated at all nodes. Further, because the best root may not be at a node, we optimize the root along the branch that gives the best test statistic, and thus find the best distance between sample 1 and 2 OTUs.

We evaluated several test statistics for the root and rate optimization (see further Appendix A), including the simple, and best performing, test statistic of summing the variances of sample (s) 1 and 2 as

$$\sum_{s=1}^{2}\sigma_s^2 = \left[\frac{1}{N_1}\sum_{i=1}^{N_1}(X_i - \overline{X}_1)\right] + \left[\frac{1}{N_2}\sum_{j=1}^{N_2}(X_j - \overline{X}_2)\right]. \qquad (1)$$

Our method can handle both unrooted and rooted Newick trees. A web version of this method is available at the Los Alamos HIV sequence database (www.hiv.lanl.gov), and is named TreeRate. The output gives $\overline{X}_1$, $\overline{X}_2$, s_1^2 and s_2^2, Σs_s^2, and $\Delta\hat{d}$ for every node in the tree and the best rooting point. The web tool also allows the user to input the time points at which each sequence was sampled, in which case the evolutionary rate, ER=$\Delta\hat{d}/\Delta t$, is also calculated for every rooting node. The time interval Δt is calculated as the arithmetic mean of the sequences with an associated time point.

Simulations

To evaluate how the rate and root optimization performed when data was limited, we tested the method under several limiting conditions, including different expected distances (Δd), fraction of the tree that contained the expected distance ($\Delta d/H$), number of taxa, sequence length, and uncertainty in the tree topology.

Random tree topologies were generated using MacClade (Maddison et al., 2003). Branch lengths were added to simulate different genetic distances from the root as well as between sample 1 and 2. Branch lengths were randomly Poisson distributed around the expected values. At distances smaller than 0.001 substitutions/site trees will become uninformed because there will be very few substitutions between taxa, and conversely at very high distances alignments become a serious limitation. Therefore, we simulated trees in a biologically typical range where the expected distance between sample 1 and 2 (Δd) ranged from 0.001 to 0.1 substitutions/site in 10 even logarithmic steps. This expected distance occurred at ratios 0.2, 0.5 and 0.8 of the total tree height ($\Delta d/H$) (Fig 1). The number of OTUs varied from 2 to 20 in sample 1 with sample 2 constant at 20, and 2 to 20 in sample 2 with sample 1 constant at 20. In all simulations the sequence length was 1000 nt, except for when the effect of sequence length was investigated, where it was varied from 100 to 100000 nt. To include uncertainty in the topology, i.e., dealing with incorrectly reconstructed trees, we generated sequences (1000 nt) using Seq-Gen (Rambaut et al., 1996), under a general-time-reversible model with Gamma distributed variation across sites according to a realistic HIV-1 situation (Leitner et al., 1997). Subsequently, a neighbor

joining (BioNJ) tree was reconstructed using PAUP* (Gascuel, 1997; Swofford, 2002) with the identical model as used to generate the sequences. Note that the tree uncertainty tests do not depend on how the trees were reconstructed; all we wanted to measure is the effect of not perfectly reconstructed trees. In all simulations 100 random trees were investigated at each setting and the root was optimized using the above test statistic (MSV). The inferred root and $\Delta\hat{d}$ were registered and compared to the true root and Δd.

Comparison to other methods

We compared the accuracy and the computing time of our method to two alternative strategies for estimating the evolutionary rates from longitudinal data (Table 1).
The mean pairwise distance (MPD), which is the fastest way to calculate genetic differences between two (or more) samples; and 2) Bayesian Markov Chain Monte Carlo (BMCMC) simulations assuming explicit clock and population growth models (Drummond et al., 2006), which is one of the perhaps most rigorous ways to estimate genetic differences. The MCMC analyses were performed using BEAST (Drummond et al., 2007), with the substitution rates generated using a general-time reversible substitution model with gamma distribution and invariable rates among sites, and Markov Chain Monte Carlo runs of 10,000,000 steps sampled every 1,000 steps and analyzed with Tracer (beast.bio.ed.ac.uk/Tracer) with a discarded burn-in of 10%. The three methods were compared using three different subtype B HIV-1 datasets consisting of U.S.A. sequences covering the V3 region. Dataset 1 contained 52 sequences sampled at two time points, 1986 and 1997, dataset 2 contained 887 sequences sampled between 1978 and 2006, and dataset

3 contained 21 sequences sampled at three different time points, 1981, 1990, and 2000.

Reconstruction of HIV trees and TreeRate analyses

HIV-1 subtype B and C phylogenies were inferred using PhyML (Guindon et al., 2003), with a general-time-reversible DNA substitution model with invariable sites and Gamma distributed variable site rates. Starting trees for the heuristic search were derived by the BioNJ method and refined by SPR and NNI improvements. Viral divergence was calculated using TreeRate by calculating Δd between sequences sampled in 1978+1979 and all other sampling times for the subtype B epidemic in Europe and North America and B epidemic in U.S.A., and between sequences sampled in 1984+1985 and all other sampling time points from the subtype C epidemic and C sub-epidemic in Ethiopia, respectively. We performed linear regression analyses of this data, and tested for the difference in slopes before and after all sampling time points using lm in R (R Development Core Team, 2003), testing for the interaction of a dummy variable "before" and "after" a possible breaking point in time showing change in the slope. The change in the slope was assessed with an indicator, $\log |s1/s2|$, where s1 is the slope "before" and s2 "after" the breakpoint, followed by a F-test for significance.

RESULTS

Identifying the optimality criterion

In the case when all branches were perfect, i.e., there was no variation in tip heights in each sample, the correct root and rate were always recovered (data not shown). Such a situation may be the case when sequences are infinitely long, but will never occur in real data. Therefore, to evaluate our method and its capacity to infer the correct genetic distance (Δd), and thus the rate of evolution, we simulated 25350 trees that aimed at limiting the information about the distance from the root to the OTUs. For the best root the distance between the two samples ($\Delta\hat{d}$) was estimated and compared to the correct genetic distance as $\Delta\hat{d}/\Delta$d.

We evaluated several test statistics to optimize the root and evolutionary rate in a given tree. Overall, the best rooting was found with the minimum sum of tip height variances (MSV) (Fig S1). This criterion performed well at low Δd, increased its rooting accuracy at higher Δd, and was not sensitive to H. The best criterion to find the optimal rate was also MSV which showed no bias to over or underestimate at any rate investigated (Fig 2 & 3).

Effect of low rates

The MSV optimality criterion showed no bias in its average estimate of the evolutionary rate at different Δd/H ratios (Fig 2). At low Δd, however, stochastic effects on branch lengths may cause individual trees to display quite a large variation and thus over- or underestimate the rate by a factor of 2 (at 0.001 substitutions/site and low Δd/H ratio). Trees reconstructed from sequences that are expected to only have moved apart 0.001 substitutions/site are not very reliable in the first place, and thus it is no surprise that the

rate may be off by a factor 2 in such cases. In fact, at this low rate we observed cases where sample 2 had evolved less than sample 1, giving negative rate values. The dispersion decreased with higher Δd and Δd/H ratios, and in general the expected error in the estimate from a single tree was less than 10% at rates when Δd>0.01 substitutions/site at all Δd/H ratios.

Effect of few taxa

With few OTUs in either sample the Δd estimation became more uncertain, but the effect of few OTUs was not as severe as one might have expected (Fig S2). At Δd=0.01 substitutions/site, only 2 OTUs in either sample caused $\Delta d/\Delta\hat{d}$ tree ratios to be off by a factor 2 or worse, but at higher rates even this sparse representation gave reasonable estimates in individual trees. There was a trend suggesting that fewer OTUs in sample 1 was worse than fewer in sample 2, explained by sample 2 having accumulated more substitutions and thus being more informative about its average height than sample 1. With more than 4 OTUs in either sample there was only slight improvement in the dispersion when more OTUs were added, and at Δd/H =0.8 even 2 OTUs gave very little variation around the average.

Effect of sequence length

Longer sequences means more information about branch lengths and less stochastic error, and thus more defined height estimates. When the part of the tree that informs about Δd is

small (Δd/H=0.2), sequence length becomes more important (Fig S3). This situation occurs when one is investigating recent events in a deep phylogeny. Hence, at Δd=0.001 substitutions/site and Δd/H=0.2 close to a sequence length of 3000 characters was required to lower the variation around $\Delta\hat{d}$ to within 10% of the true rate. At higher Δd and Δd/H ratios the precision got much better. Many biological studies involve sequence lengths in the 300-10000 range (average length in GenBank is approximately 1000 nt (Benson et al., 2007)), and at the lower end of this range (l=300-1000) our $\Delta\hat{d}$ estimates had good precision (var [$\Delta\hat{d}/\Delta$d]<1.0) at all Δd's for Δd/H =0.8 and at roughly Δd>0.0063 substitutions/site for Δd/H ≥0.2).

Effect of uncertain tree topology

To assess the case when we do not have the correct tree, but rather a reasonable tree, we investigated trees that were reconstructed from DNA sequence data generated on random trees with 20 OTUs in each of two longitudinal samples. There was a clear correlation between the accuracy of the tree reconstruction and Δd, i.e., at low Δd the trees were less accurately reconstructed (Fig 3). As expected, finding an accurate rate was easier at higher expected rates. In general, at Δd>0.003 substitutions/site the estimated rate was within 10% of the true rate, regardless of how inaccurate the reconstructed tree was. Interestingly, at higher Δd/H ratios the trees were more inaccurate, because H was smaller, but the estimated rates were still good. Thus, the rate estimation was robust to errors in the (topological) tree reconstruction, which is important for real situations.

Finding the correct root

Strictly, finding the correct root requires the true tree to be recovered. Thus, we investigated the probability of finding the correct root given the true tree. The likelihood of finding the correct root increased with higher Δd, sequence length and number of OTUs in samples 1 and 2, but decreased with higher Δd/H ratios. At Δd/H=0.2, the success of finding the correct root was 24% at Δd=0.001 substitutions/site, then increased to 83% at Δd=0.1 substitutions/site, while at Δd/H =0.8 the success went from 6 to 26% (Fig S2). Similarly, increased sequence length had a stronger positive effect on the success of finding the correct root when Δd/H was low. Finally, when there were limitations in the number of OTUs in either sample (N<20), the root was more often found in the correct location when $\Delta\hat{d}$ was high.

Comparison to other methods: TreeRate is both accurate and fast

We compared the accuracy and the computing time of our method to two alternative strategies for estimating the evolutionary rates from longitudinal data, MPD and BMCMC (Table 1). As expected, the MPD was very fast and the BMCMC very slow. In scientific context accuracy is often more important than speed, however. Thus, considering accuracy in the evolutionary rate estimation first and speed second, our method gave accurate estimates at fast calculation speed, even when including tree building using ML (our method was >200 times faster than BMCMC on a 52 taxa set). The MPD gave inaccurate rates; with no homoplasy it is expected to seriously overestimate the differences, but with

HIV data which has a high degree of homoplasy the overestimation will be gradually compensated for so that with larger number of taxa the differences become underestimated (Table 1). Thus, our method is well suited for analysis of large and numerous datasets.

Application to real HIV-1 data

We collected HIV-1 DNA sequences that covered the *env* V3 region with at least 324 and 285 nt in the HIV database (hiv.lanl.gov) from the subtype B and C epidemics, respectively (B, 887 and C, 744 sequences). We confirmed that the sequences came from the same general respective epidemic by reconstruction of large phylogenetic trees (data not shown). For instance, only subtype C sequences from the African C epidemic were included and not Indian C which form a distinct cluster, indicating a separate epidemic. Similarly, subtype B sequences from North America and Europe were confirmed to belong to the same epidemic.

Figure 4 shows the real variances σ_1^2 and σ_2^2 that our root optimization is based on (MSV) compared to the expected Poisson variances for the optimized heights $\overline{X}_1$ and $\overline{X}_2$ of the subtype B data. Three observations justify the assumption of a fairly constant rate in each time interval: 1) The real variances were proportional to the expected Poisson variances ($R^2 \approx 0.76$). 2) As $\overline{X}_1$ and $\overline{X}_2$ grew over time, so did σ_1^2 and σ_2^2, suggesting a Poisson process. 3) Samples from time point two generally had larger variance than those of time point one in each comparison ($p<0.01$, t test), which would be expected if $\overline{X}_1$ and $\overline{X}_2 \sim$ Pois(l_i) and $\overline{X}_1 < \overline{X}_2$. Note that the assumption of a constant rate only applies to each

investigated time interval, and that this makes it possible to find rate changes over time, as we show below. This also allows to test at which time interval a constant rate is robustly inferred (here that was at $\Delta t \geq 3$years).

Subtype B and C epidemics display complex evolutionary rates

Both subtype B and C displayed evolutionary rates with relatively large fluctuations over time (Fig S5). When comparing our results to HIV-1 prevalence data (www.unaids.org), it became clear that both epidemics consisted of sub-epidemics with different dynamics in the countries involved, i.e., while the prevalence increased in one country the prevalence went down in another. Thus, the uneven sampling from sub-epidemics that progress with different dynamics may explain a large portion of the fluctuations. Subtype C showed larger fluctuations over time than subtype B, agreeing with the fact that the epidemic dynamics in African countries are much more diverse than those in European and North American countries.

Dynamics in an epidemic are reflected in the evolutionary rate

To decipher the complex overall pattern of the larger subtype B and C epidemics, we analyzed the two countries we had most data from; U.S.A. (subtype B) and Ethiopia (subtype C). The HIV-1 subtype B epidemic in the U.S.A. showed a significant decrease ($p<0.001$, F-test) in the rate of evolution after 1997 (Fig 5A). Interestingly, while the prevalence kept stable at 0.6% this change in the rate of evolution coincided with the onset of HAART in the U.S.A. (and Europe) in 1996. The overall subtype B epidemic in North America and Europe showed the same result ($p<0.001$, F-test). The subtype C epidemic in

Ethiopia had a clear stagnation in prevalence around 1995-1996 (Fig 5B). While we had much more limited longitudinal sequence data available for this epidemic, the decrease in the epidemic rate was tracked by an increase in the evolutionary rate ($p>0.05$, F-test). Thus, when the epidemic rate changes, then the evolutionary rate of the virus inversely reflects that in a dynamic way. These results indicate that a sudden change in an epidemic may be reflected in the rate of evolution of the virus on the population level.

DISCUSSION

Large DNA sequence datasets with longitudinal samples have become common, especially for rapidly evolving organisms such as HIV. With the recent development of ultra-high throughput sequencing these already large datasets will become even larger. Large datasets from epidemics may inform about the rate of spread, and thus signal about outbreaks and other changes in the epidemic. Since our method is both fast and accurate, it may be used to efficiently analyze such data.

We used TreeRate to assess the evolutionary rate and epidemiological history of HIV-1 subtypes B and C. It has previously been suggested that there are subtype-specific differences in the patterns of epidemic growth of subtypes B and C (Walker et al., 2005). Our results showed that the evolutionary rate of both subtypes displayed relatively large fluctuations over time, with subtype C having larger fluctuations than subtype B, agreeing with the fact that the epidemic dynamics in African countries are much more diverse than

those in European and North American countries. When compared to HIV-1 prevalence data from countries that the samples for subtype C were derived from, it became clear that this epidemic consisted of several sub-epidemics with different dynamics, explaining the fluctuations in the evolutionary rate over time.

Thus, we investigated the evolutionary rates for two sub-epidemics from countries we had most data from: Ethiopia for subtype C, and U.S.A. for subtype B. In Ethiopia, subtype C is the most dominating subtype, and the introduction of HIV-1 into this country has been estimated to 1983 (1980-1984) (Abebe et al., 2001). By analyzing the divergence of HIV-1 from 1984+1985 (the earliest available sequences in the LANL HIV database) to all subsequent sampling time points up to 2005, we observed an indication of a dynamic inverse correlation between virus spread and the evolutionary rate. Prevalence data from Ethiopia show that HIV-1 prevalence increased until about 1995, from which point it started to slowly decrease. Although the change in the slope was borderline significant, likely due to sparse data, this trend indicates that it is possible to study epidemic dynamics by consecutively estimating the rate of evolution of HIV-1 on the population level.

For subtype B, there was a significant decrease in the rate of evolution at the time of introduction of HAART in U.S.A. (and Europe). If antiretroviral therapy is successful, the viral replication within a host will be diminished, and there would be no measurable accumulation of substitutions in *env*. It has previously been shown that effective antiretroviral treatment can slow down and even totally abolish the evolution of HIV-1 in the envelope region (Drummond et al., 2001; Nijhuis et al., 1998; Rodrigo et al., 2003). It is

possible that this effect is reflected in the decrease of the evolutionary rate of subtype B on the population level. However, it is also possible that HAART effectively diminishes the number of HIV-1 transmissions in the chronic stage of infection due to successful reduction of viral load, thus skewing the transmissions of the virus to the acute phase of infection. We have previously shown that the rate of evolution of HIV-1 is lower if it is spread rapidly in a population, when most of the individuals are still in the acute phase of infection, before the HIV-1-specific immune system has a chance to exert pressure on the virus to change (Maljkovic Berry et al., 2007). The exact mechanism of successful antiretroviral treatment on the rate of evolution of HIV-1 needs to be further evaluated, as the use of HAART is increasing throughout the world and will affect other subtypes than B. By studying the effect of HAART on subtype B we might thus be able to predict the effect of HAART on the HIV-1 pandemic as a whole.

Several studies have indicated that HIV-1 subtype B had spread rapidly in the initial stages of the epidemic in the U.S.A. (Gilbert et al., 2007; Robbins et al., 2003; Selik et al., 1984; Walker et al., 2005), with a slow-down of the rate of new infections in the beginning of the 1990s. With this data, we would expect to see a lower evolutionary rate of subtype B before 1990. This trend is not observed in our analysis, agreeing with a suggestion that the period of exponential growth of US subtype B precedes most of the early documented cases (Robbins et al., 2003). Introduction of HIV-1 subtype B into the US has been estimated to have occurred in or around 1969 (1962-1970) (Gilbert et al., 2007). This suggests that the virus circulated in the country for about 12 years before recognition of AIDS in 1981. Since there are essentially no HIV sequences for this period, it is impossible to tell how fast the

virus was spreading in the US population during this time. However, data on increase of STDs and other rare infections among men who have sex with men (MSM), the risk group initially affected by HIV subtype B in the U.S.A., suggest that the virus might have been spreading rapidly during this silent period. For instance, in the MSM risk group, between 1974 and 1979 amebiasis cutis ulcers increased by 250%, hepatitis A case reports doubled, and hepatitis B cases tripled (Garrett, 1995). In 1981, a study was published showing that the number of active cytomegalovirus (CMV) cases jumped in less then a decade from 10% to over 94% among MSM (Drew et al., 1981). CMV has been associated with AIDS since the first reports of the epidemic in the MSM risk group. Thus, although it is possible that HIV-1 spread rapidly in the initial silent phase of the epidemic, our results indicate that the rate of spread had slowed down by the time of sampling of first HIV-1 sequences.

It is well known that HIV recombines during its evolution (Leitner et al., 1995; Robertson et al., 1995; Sabino et al., 1994). If recombination occurs in phylogenetic trees, this undermines the fundamental assumption of a binary structure, and thus topology and branch lengths may become inaccurate. However, it is possible that HIV-1 recombination may have a larger effect on the population level. In fast spread of the virus, such as in standing social IDU networks, the chances of superinfection, and thus recombination, are greater, suggesting that fast epidemics may have a higher rate of virus recombination. This may affect the assessment of the evolutionary rate on the population level, and is something that should be analyzed in the future, and is out of scope for this paper. Furthermore, it is unlikely that the amount of recombination will drastically change during an individual epidemic such as in our analyses of subtypes B and C over time, making recombination a

contributing but constant factor in these analyses.

The HIV trees were inferred using a maximum likelihood method with no assumption of a molecular clock, i.e., all branches were free to vary. Thus, the variance we estimate will inform how "clocklike" a tree is. A fairly strict clock is likely to hold for closely related species or, as the primary intent of our method, for within-population estimates (Kishino et al., 1990; Rambaut et al., 1998; Yoder et al., 2000). In the HIV data investigated here, we found that the rate in one time interval can follow a Poisson distributed clock quite well (Fig 4), but that temporal changes in the evolutionary rate may occur as the result of epidemic dynamics (Fig 5).

Although we were able to find the correct root in 100% of our simulations when the sequence length was very high (100,000 nt) and $\Delta d > 0.006$ substitutions/site at $\Delta d/H = 0.2$, it appeared that our method in general was not very efficient at finding the correct root. This is not surprising because there will be very few, if any, substitutions on expected short branches, making it impossible to resolve the whole tree and thus to find the true topology and the correct root (e.g., Fig 1C). In spite of this, the rate estimates were generally good, within 10% of the true rate. This happens because when there are no or very few substitutions on expected short branches close to the true root, it does not matter from which exact topological point on the tree one estimates $\overline{X}_1$ and $\overline{X}_2$, such short branches may mislead the exact rooting but not the overall evolutionary rate.

In a real situation, when we reconstruct a phylogeny from sequence data, we may never know if we have found the true tree, and thus the true root may be impossible to find. It is well known that tree reconstruction and rooting is especially difficult in cases where there is a combination of short and long branches. This may be due to the effect of long branch attraction (Bruno et al., 1999; Felsenstein, 1978) to misspecification of the substitution model (Ho et al., 2004; Kolaczkowski et al., 2004; Mar et al., 2005), or to limitations of the heuristic used to explore alternative branching patterns. Similarly, rooting has been shown to be particularly difficult in trees displaying rapid radiations (Shavit et al., 2007). Thus, in addition to when there is too little information on some branches to resolve the tree, in real situations when trees are reconstructed, topologies, branch lengths and roots may also be mislead due to methodological artifacts and inaccurate substitution models. Importantly, our method was robust to inaccurately reconstructed trees (Fig 3). The simulated trees were reconstructed using NJ, and it is possible that our Δd estimates would have been even better if we had used ML (as in the HIV inferences) to reconstruct the topology and, in this context more importantly, the branch lengths.

Estimating root-to-tip distances from a non-star tree does not give independent data (Felsenstein, 1985; Felsenstein, 2004), and thus this may bias the true variances of the distances in the samples. This is because branches deeper into the tree are reused and can influence several root-to-tip distances up or down. In comparative studies it has been clearly shown that hierarchically structured phylogenies create statistical problems if traits of the taxa under study are treated as if drawn independently from the same distribution,

e.g., (Dessimoz et al., 2008; Felsenstein, 1985; Ives et al., 2007; Kelly et al., 2004; Symonds, 2002). For instance, the resulting covariance can be taken into account using the method of generalized least squares (GLS) while ordinary and weighted least squares methods (OLS and WLS), such as the well-known Fitch-Margoliash method (Felsenstein, 1997; Fitch et al., 1967) implemented in for instance PHYLIP and PAUP (Felsenstein, 1993; Swofford, 2002), assume independent distance estimates. However, both OLS and GLS based methods yield unbiased estimates of regression coefficients (Pagel, 1993), and interestingly the deviations from OLS have been shown to be greater than from GLS, i.e., the variance was overestimated rather than underestimated when non-independence was not accounted for (Rohlf, 2006). Importantly, just as OLS is not biased, though less efficient than WLS and GLS, our rate estimation method does not systematically bias the choice of root. In any case, we find that when the root is incorrectly estimated, our rate estimate is still good and unbiased.

In conclusion, we have evaluated a simple method that optimizes the root and evolutionary rate in a given tree. The taxa in the tree must have at least two timestamps and realistic branch lengths. The two samples of taxa can, for instance, come from two samples of a population separated by a time interval, but not divided into separate monophyletic groups. We have shown that this method performs well in estimating the evolutionary rate under a large interval of expected rates, sequence lengths, and limited number of taxa. The method was less efficient in finding the true root, but the evolutionary rate estimation was robust against rooting errors and inaccuracies in the tree topology. Applied to real HIV-1 data, we found that when changes occur in an epidemic, such as changes in the rate of spread of the

virus, or introduction of effective antiretroviral treatment, then the evolutionary rate of HIV-1 at the population level reflects these changes. In addition, we show that the rate of evolution of HIV-1 can differ in different stages of an epidemic, which may have implications on the estimations of the most recent common ancestor and the time of introduction of HIV-1 in a population. Thus, it is possible that the estimations on the time of introduction of HIV-1 into *Homo Sapiens* may have to be re-evaluated.

ACKNOWLEDGEMENTS

This study was funded by a NIH/DOE interagency agreement (A1-YI-1500) and approved by LANL (LA-UR 08-0806). We thank Catherine Macken and Sydeaka Watson for helpful discussions and technical assistance with the statistics of our analyses.

APPENDIX A. Alternative optimality criteria

In this paper we evaluated 7 test statistics for the root and rate optimization (Fig S1). The four best criteria to find the true root were minimizing the sum of the tip height variances of OTUs in both samples as in Eq. 1 (MSV), maximizing Welch's t-value, minimizing Welch's p-value (MWP) (Welch, 1947), and minimizing either of the two samples' variance. For Welch's t test, the t statistic is calculated as

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

where $\overline{X}_i$ is the mean distance to the root of sample i, σ_i^2 the sample variance, and N_i the sample size. Thus, this allows for unequal variances in sample 1 and 2. To calculate the p-value for each root, the degrees of freedom v were estimated as

$$v = \frac{\left(\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}\right)^2}{\frac{\sigma_1^4}{N_1^2 \bullet v_1} + \frac{\sigma_2^4}{N_2^2 \bullet v_2}}$$

where v_i is the degrees of freedom associated with the i^{th} variance estimate N_i-1. The p-value calculations were done using R (R Development Core Team, 2003). While MWP performed well at higher Δd and Δd/H ratios, it was sensitive to total tree height (H). The t test statistic (MWP) had a bias at low Δd, while our $\Delta\hat{d}$ estimates were unbiased across all rates using MSV (Fig S1 and Fig 2). We compared MSV and MWP to the upper and lower boundaries (maximizing and minimizing $\Delta\hat{d}$, respectively), to minimizing either sample's variance, and to the theoretical limit of our simulations, i.e., the rate estimated at the true root. As we have noted previously, MWP overestimated $\Delta\hat{d}$ when Δd was below 0.003 substitutions/site (Maljkovic Berry et al., 2007). While this is a very low rate, with only 3 substitutions on average in a 1000 nt long sequence, MSV showed no bias even at very low rates (Fig S1). In conclusion, MSV was found to be the best optimality criterion for finding the true root and rate in a given tree.

The difference between two Poisson distributed variables is skewed according to the Skellam distribution (Skellam, 1946). Qualitatively, this skewness has the same behavior as the MWP bias, i.e., more positive bias at lower Δd, but quantitatively it had an effect 50-fold below what we observed. Thus, although the Skellam skewness is in effect, it drowns in the phylogenetic noise and has no practical effect on our $\Delta\hat{d}$ estimates. Interestingly, some obscure criteria performed well for specialized conditions, e.g., minimizing the average tip height to sample 1 OTUs displayed overall high performance maxima that depended on the relationship of Δd and H (data not shown), but using this for general purposes would be unpractical unless one knew what to expect and was able to collect samples in an optimal way. Also interesting to note was that neither minimizing nor maximizing $\Delta\hat{d}$ ever found the correct root (Fig S1).

APPENDIX B. Supplementary results.

Supplementary data associated with this article can be found in the online version at doi:.

REFERENCES

Abebe, A., Lukashov, V., Pollakis, G., Kliphuis, A., Fontanet, A., Goudsmit, J. and de Wit, T. 2001. Timing of the HIV-1 subtype C epidemic in Ethiopia based on early virus strains and subsequent virus diversification. AIDS 15 1555-1561.

Aris-Brosou, S. 2007. Dating phylogenies with hybrid local molecular clocks. PLoS ONE 2 e879.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Wheeler, D. L. 2007. GenBank. Nucleic Acids Res 35 D21-5.

Bruno, W. J. and Halpern, A. L. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol Biol Evol 16 564-6.

Dessimoz, C. and Gil, M. 2008. Covariance of maximum likelihood evolutionary distances between sequences aligned pairwise. BMC Evol Biol 8 179.

Drew, W. L., Mintz, L., Miner, R. C. and Ketterer, B. 1981. Prevalence of cytomegalovirus infection in homosexual men. J Infect Dis 143 188-192.

Drummond, A., Forsberg, R. and Rodrigo, A. 2001. The inference of stepwise changes in substitution rates using serial sequence samples. Mol Biol Evol 18 1365-1371.

Drummond, A., Rambaut, A., Shapiro, B. and Pybus, O. G. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22 1185-1192.

Drummond, A. J., Ho, S. Y., Phillips, M. J. and Rambaut, A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol 4 e88.

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

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Systematic Zoology 27 401-410.

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125 1-15.

Felsenstein, J. (1993). PHYLIP: Phylogeny Inference Package. Seattle, WA, University of Washington.

Felsenstein, J. 1997. An alternating least squares approach to inferring phylogenies from pairwise distances. Syst Biol 46 101-11.

Felsenstein, J. (2004). Inferring phylogenies. Sunderland, MA, Sinauer Assocates.

Fitch, W. M. and Margoliash, E. 1967. Construction of phylogenetic trees. Science 155 279-284.

Garrett, L. (1995). The Coming Plague, Penguin Group (USA) Inc., 375 Hudson Street, New York, New York 10014, USA.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14 685-95.

Gilbert, M. T., Rambaut, A., Wlasiuk, G., Spira, T. J., Pitchenik, A. E. and Worobey, M. 2007. The emergence of HIV/AIDS in the Americas and beyond. Proceedings of the National Academy of Sciences USA 104 18566-18570.

Gillespie, J. H. 1984. The molecular clock may be an episodic clock. Proceedings of the National Academy of Sciences USA 81 8009-8013.

Gillespie, J. H. 1988. More on the overdispersed molecular clock. Genetics 118 385-388.

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

Ho, S. Y. and Jermiin, L. 2004. Tracing the decay of the historical signal in biological sequence data. Syst Biol 53 623-37.

Huelsenbeck, J. P., Larget, B. and Swofford, D. 2000. A compound poisson process for relaxing the molecular clock. Genetics 154 1879-92.

Ives, A. R., Midford, P. E. and Garland, T., Jr. 2007. Within-species variation and measurement error in phylogenetic comparative methods. Syst Biol 56 252-70.

Kelly, C. and Price, T. D. 2004. Comparative methods based on species mean values. Math Biosci 187 135-54.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16 111-120.

Kishino, H. and Hasegawa, M. 1990. Converting distance to time: application to human evolution. Methods Enzymol 183 550-70.

Kishino, H., Thorne, J. L. and Bruno, W. J. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol Biol Evol 18 352-61.

Kolaczkowski, B. and Thornton, J. W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431 980-4.

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

Leitner, T. and Albert, J. 1999. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proceedings of the National Academy of Sciences USA 96 10752-10757.

Leitner, T., Escanilla, D., Marquina, S., Wahlberg, J., Brostrom, C., Hansson, H. B., Uhlen, M. and Albert, J. 1995. Biological and molecular characterization of subtype D, G, and A/D recombinant HIV-1 transmissions in Sweden. Virology 209 136-146.

Leitner, T., Kumar, S. and Albert, J. 1997. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. Journal of Virology 71 4761-4770 (see also correction 1998: 72; 2565).

Maddison, D. R. and Maddison, W. P. (2003). MacClade 4: Analysis of Phylogeny and Character Evolution. Sunderland, MA, Sinauer.

Maljkovic Berry, I., Ribeiro, R., Kothari, M., Athreya, G., Daniels, M., Lee, H. Y., Bruno, W. and Leitner, T. 2007. Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. J Virol 81 10625-35.

Mar, J. C., Harlow, T. J. and Ragan, M. A. 2005. Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. BMC Evol Biol 5 8.

Nijhuis, M., Boucher, C., Schipper, P., Leitner, T., Schuurman, R. and Albert, J. 1998. Stochastic processes strongly influence HIV-1 evolution during suboptimal protease-inhibitor therapy. Proceedings of the National Academy of Sciences USA 95 14441-14446.

Pagel, M. 1993. Seeking the evolutionary regression coefficient: an analysis of what comparative methods measure. J Theor Biol 164 191-205.

R Development Core Team (2003). R: A language and environment for statistical computing., R Foundation for statistical computing, Vienna, Austria.

Rambaut, A. and Bromham, L. 1998. Estimating divergence dates from molecular sequences. Mol Biol Evol 15 442-8.

Rambaut, A. and Grassly, N. (1996). Sequence-Generator: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Oxford, University of Oxford.

Robbins, K. E., Lemey, P., Pybus, O. G., Jaffe, H. W., Youngpairoj, A. S., Brown, T. M., Salemi, M., Vandamme, A. M. and Kalish, M. L. 2003. U.S. Human immunodeficiency

virus type 1 epidemic: date of origin, population history, and characterization of early strains. J Virol 77 6359-6366.

Robertson, D. L., Sharp, P. M., McCutchan, F. E. and Hahn, B. H. 1995. Recombination in HIV-1. Nature 374 124-126.

Rodrigo, A. G., Goode, M., Forsberg, R., Ross, H. A. and Drummond, A. 2003. Inferring evolutionary rates using serially sampled sequences from several populations. Mol Biol Evol 20 2010-8.

Rohlf, F. J. 2006. A comment on phylogenetic correction. Evolution 60 1509-15.

Sabino, E. C., Shpaer, E. G., Morgado, M. G., Korber, B. T., Diaz, R. S., Bongertz, V., Cavalcante, S., Galvao-Castro, B., Mullins, J. I. and Mayer, A. 1994. Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from Brazil. J Virol 68 6340-6346.

Salemi, M., Strimmer, K., Hall, W. W., Duffy, M., Delaporte, E., Mboup, S., Peeters, M. and Vandamme, A.-M. 2001. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. FASEB Journal 15 276-278.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol 19 101-9.

Selik, R. M., Haverkos, H. W. and Curran, J. W. 1984. Acquired immune deficiency syndrome (AIDS) trends in the United States, 1978-1982. Am J Med. 76 493-500.

Shavit, L., Penny, D., Hendy, M. D. and Holland, B. R. 2007. The problem of rooting rapid radiations. Mol Biol Evol 24 2400-11.

Skellam, J. 1946. The frequency distribution of the difference between two Poisson variates belonging to different populations. Journal of the Royal Statistical Society: Series A 109 296.

Swofford, D. L. (2002). PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods). Sunderland, MA, Sinauer Associates.

Symonds, M. R. 2002. The effects of topological inaccuracy in evolutionary trees on the phylogenetic comparative method of independent contrasts. Syst Biol 51 541-53.

Takahata, N. 1987. On the overdispersed molecular clock. Genetics 116 169-179.

Thorne, J. L., Kishino, H. and Painter, I. S. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15 1647-57.

Walker, P. R., Pybus, O. G., Rambaut, A. and Holmes, E. C. 2005. Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth. Infect Genet Evol. 5 199-208.

Welch, B. L. 1947. The generalization of "student's" problem when several different population variances are involved. Biometrika 34 28-35.

Yang, Z. and Rannala, B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol Biol Evol 23 212-26.

Yoder, A. D. and Yang, Z. 2000. Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol 17 1081-90.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. Evolving genes and proteins. V. Bryson and H. J. Vogel. New York, Academic Press: 97-166.

FIGURE LEGENDS

Figure 1. Definitions and examples of simulated trees. (**A**) An example of a randomly generated true tree, with perfect "clocklike" edges. H is the total tree height, and Δd is the true (expected) rate between sample 1 and 2 OTUs. This tree is at Δd/H=0.2 and 20 OTUs in each sample. Thus, this tree shows the definitions of Δd and H, and is the true tree on which the trees in panels B and C were simulated, allowing for comparison between estimated rate and expected rate ($\Delta\hat{d}/\Delta$d). (**B**) The same tree topology with Poisson distributed edges, and scaled so that Δd = 0.1 substitutions/site. $\overline{X}_1$ is the average distance from the root to sample 1 OTUs, $\overline{X}_2$ is the average distance from the root to sample 2 OTUs, and $\Delta\hat{d}$ is the estimated rate between the samples. (**C**) The same tree topology with

Poisson distributed edges, and scaled so that $\Delta d = 0.001$ substitutions/site. Note that many expected short edges become zero at this low rate, and samples 1 and 2 are not well separated. Open squares are sample 1 OTUs and filled squares sample 2 OTUs. Trees in **B** and **C** are examples of trees used in evaluating our method, scaled to the shown scale bars. The tree in **A** is of arbitrary length.

Figure 2. Estimation of $\Delta\hat{d}$ as a function of Δd.

The dashed line indicates perfect estimation of $\Delta\hat{d}$, and colored lines show the average estimates of the MSV optimality criterion, simulated at 20 OTUs in each sample and at different $\Delta d/H$ ratios. Open circles show the results from individual random trees (100 at each rate and $\Delta d/H$ ratio).

Figure 3. Estimation of $\Delta\hat{d}$ when the tree is uncertain.

The level of uncertainty, i.e., our inability to find the true tree, was measured as symmetric tree-to-tree distances (y-axis), at 11 evenly logarithmic distributed expected rates (Δd; x-axis). The estimated rate was compared to the true rate (in the true tree) and the average $\Delta\hat{d}/\Delta d$ is indicated by the color scale at the right. The resulting heat maps are at $\Delta d/H=0.2$ in **A**, $\Delta d/H=0.5$ in **B**, and $\Delta d/H=0.8$ in **C**. Each data point (colored block) is the average of 100 random simulated and reconstructed trees with 20 OTUs in each sample.

Figure 4. Comparison of HIV rate variance to Poisson variance.

The lines show the real variances σ_1^2 (blue) and σ_2^2 (red) that our root optimization was

based on compared to expected Poisson variances [$\sigma^2_{Pois(1)}$ (light blue) and $\sigma^2_{Pois(2)}$(orange)] for the optimized real heights $\overline{X}_1$ and $\overline{X}_2$ of the HIV-1 subtype B data at Δt=6 years. The expected Poisson variances were calculated from 1000 Monte Carlo simulated $X_i \sim$ Pois($\lambda_1 = \overline{X}_1$) and $X_j \sim$ Pois($\lambda_2 = \overline{X}_2$) per year (44,000 simulated root-to-tip heights). The real variances were proportional to the expected Poisson variances (scale factors 75 and 79 for samples 1 and 2, respectively).

Figure 5. Tracking the dynamics of HIV-1 epidemics.

The change in the evolutionary rate of HIV-1 on the population level (genetic divergence) dynamically tracked changes in the epidemics of subtype B in the U.S.A. (**A**) and subtype C in Ethiopia (**B**). While the prevalence was stable in the U.S.A., the change in the HIV-1 evolutionary rate coincided with the onset of HAART. In Ethiopia a change in the HIV-1 evolutionary rate indicated a dramatic change in the prevalence. An indicator variable (log |s1/s2|, where s1 is the slope before the change and s2 after the change) was used to find the best breakpoint in the evolutionary rate trend, followed by a formal F-test. The best breakpoint is shown by the dashed line. Note that the indicator has a positive value when the slope changes to a less steep value, and negative when it becomes steeper after the breakpoint. All possible breakpoints were evaluated and at least 3 divergence data points were required to calculate a slope. The resulting slopes before and after the breakpoint are plotted in the divergence graph (in **A**, s1= 0.004 and s2= 0.00001; and in **B**, s1=-0.0001 and s2= 0.01 substitutions site^{-1} year^{-1}). Each divergence data point indicates the evolutionary rate calculated from a separate tree optimized by TreeRate. The divergence in both

epidemics was calculated from the earliest available sequence samples, 1978+1979 for subtype B in the U.S.A. and 1984+1985 for subtype C in Ethiopia.

Table 1. Comparison of results and computing time to other methods

Dataset[a]	MPD[b]		BEASTcc[c]		BEASTlnsky[d]		TreeRate[e]		
	ER	t_{cpu}	ER	t_{cpu}	ER	t_{cpu}	ER	t_{cpu}	t_{cpu}Tot[f]
3s21taxa	5.97	0.012s	4.82	2.72h	3.96	3.39h	3.37	34.94s	1.06m
2s52taxa	1.22	0.036s	6.42[g]	4.79h	5.09[g]	6.76h	4.70	1.51m	1.94m
29s887taxa	1.14	7.456s	[h]		[h]		4.94	1.63m	1.92m

TABLE FOOTNOTES

ER, Evolutionary rate [10^{-3} substitutions site^{-1} year^{-1}]. t_{cpu}, computer time for actual calculations [s, seconds; m, minutes; h, hours], not considering pre- and post-processing of data (which varies but is roughly similar for all methods). All calculations were done on a computer with dual dual-core (4 CPUs) Intel® Xeon™3.20GHz CPUs with 4149768 kB memory running CentOS 5.2.

[a] The datasets consisted of HIV-1 subtype B *env* V3 region sequences (3s21taxa=426nt, 2s52taxa=282nt, 29s887taxa=324nt). The number of longitudinal samples is indicated before the "s" and the number of OTUs before "taxa". The ER was calculated between time points 1981 and 2000 for MPD and TreeRate with dataset 3s21taxa and 1985-1999 for MPD and TreeRate with dataset 29s887taxa; for 2s52taxa all OTU data was used by all methods; and BEAST used all time points available in each dataset. These datasets are available upon request from the authors; 3s21taxa is also the "Sample Input" on our web interface.

[b] MPD, mean pairwise differences among relevant OTUs calculated using PAUP* (Swofford, 2002). Distances were calculated using a general-time-reversible model with invariable sites and gamma distributed variable sites (GTR-IG).

[c] BEASTcc, BEAST estimate using default parameters with a constant clock and constant population size (Drummond et al., 2007) with a GTR-IG substitution model.

[d] BEASTlnsky, BEAST estimate using default parameters with a lognormal distributed relaxed clock and a skyline coalescent population growth model (Drummond et al., 2005) with a GTR-IG substitution model.

[e] TreeRate, the method described in this paper.

[f] t_{cpu}Tot, the total CPU time for calculating a PhyML tree with a GTR-IG substitution model (Guindon et al., 2003) plus the TreeRate root and ER optimization.

[g] These values came from BEAST runs with effective sample size <20.

[h] We were not able to get these runs started, possibly due to the large data file.

Figure 1

Maljkovic Berry et al Fig 1

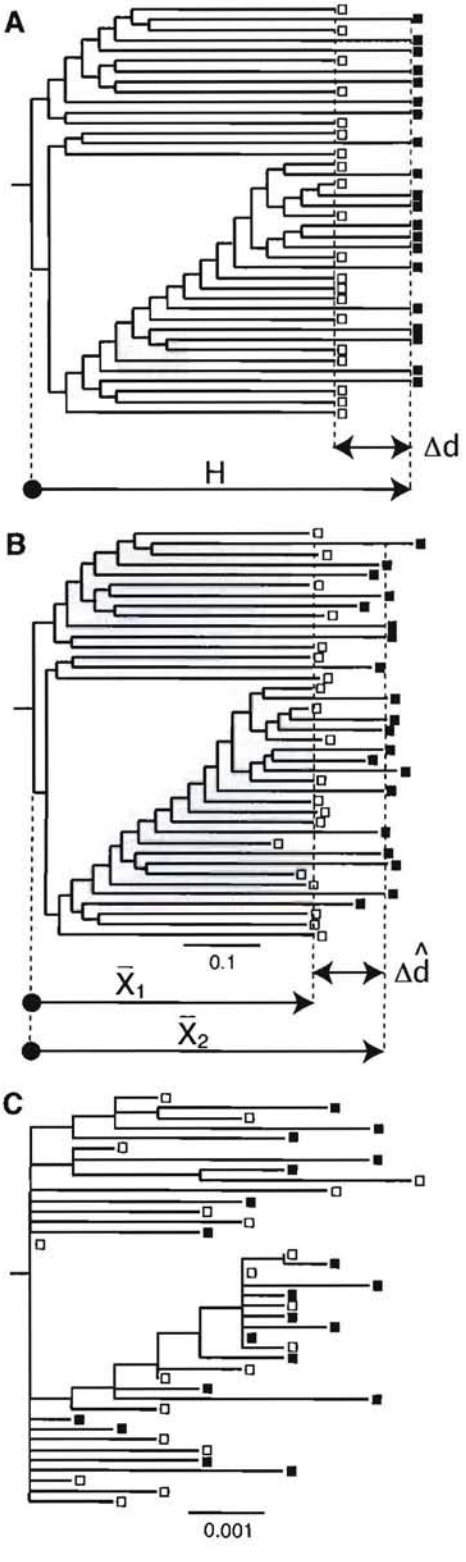


Figure 2

Maljkovic Berry et al Fig 2

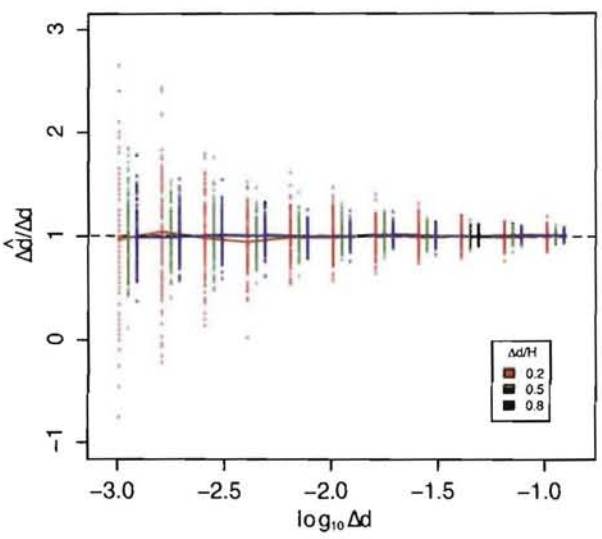


Figure 3

Maljkovic Berry et al Fig 3

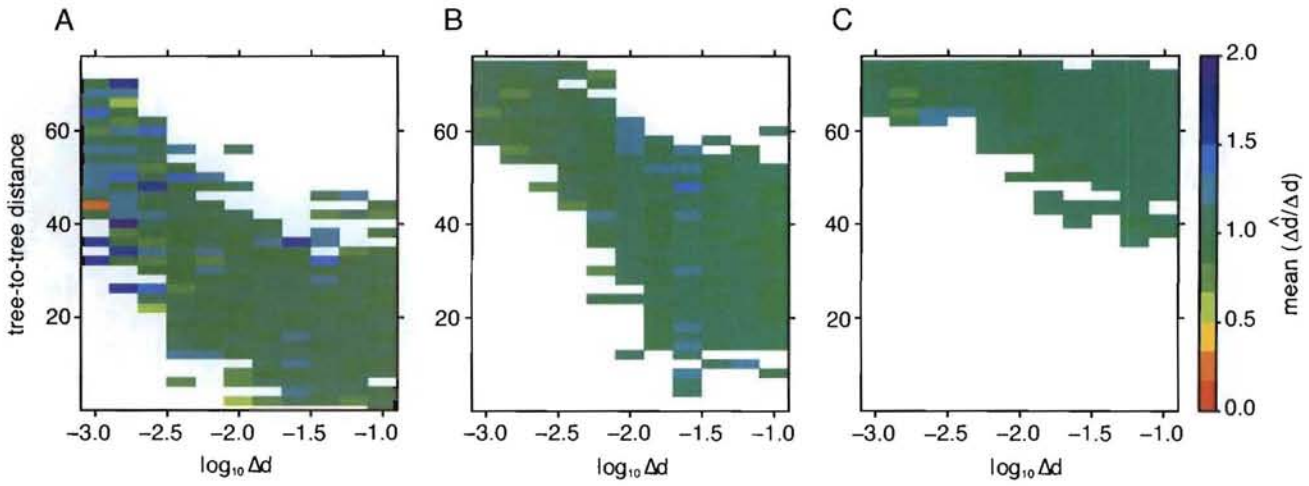


Figure 4

Maljkovic Berry et al Fig 4

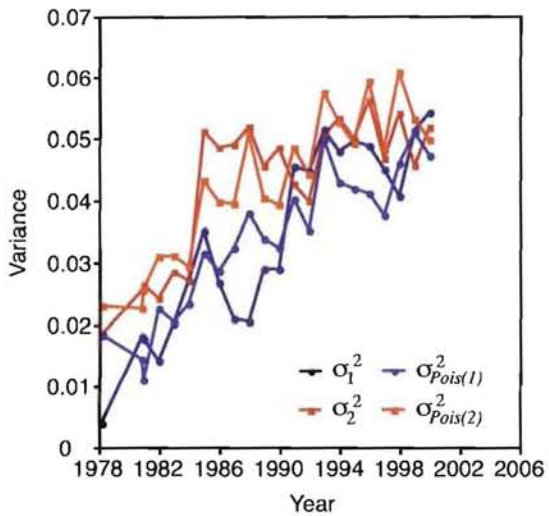
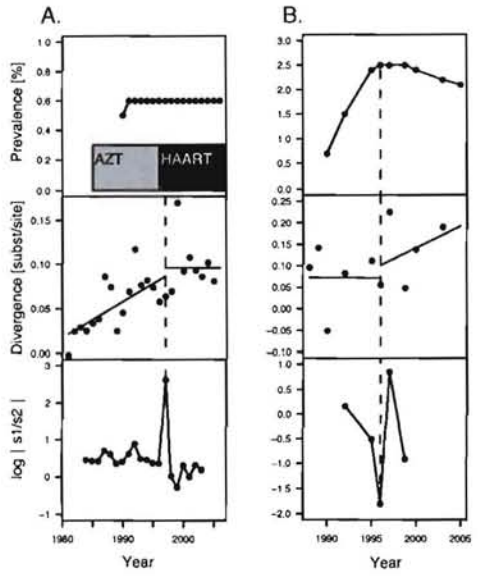


Figure 5

Maljkovic Berry et al Fig 5



The evolutionary rate dynamically tracks changes in HIV-1 epidemics: application of a simple method for optimizing the root in phylogenetic trees with longitudinal data

Irina Maljkovic Berry, Gayathri Athreya, Moulik Kothari, Marcus Daniels, William J. Bruno, Bette Korber, Carla Kuiken[a] Ruy M. Ribeiro, Thomas Leitner

APPENDIX B. SUPPLEMENTARY RESULTS

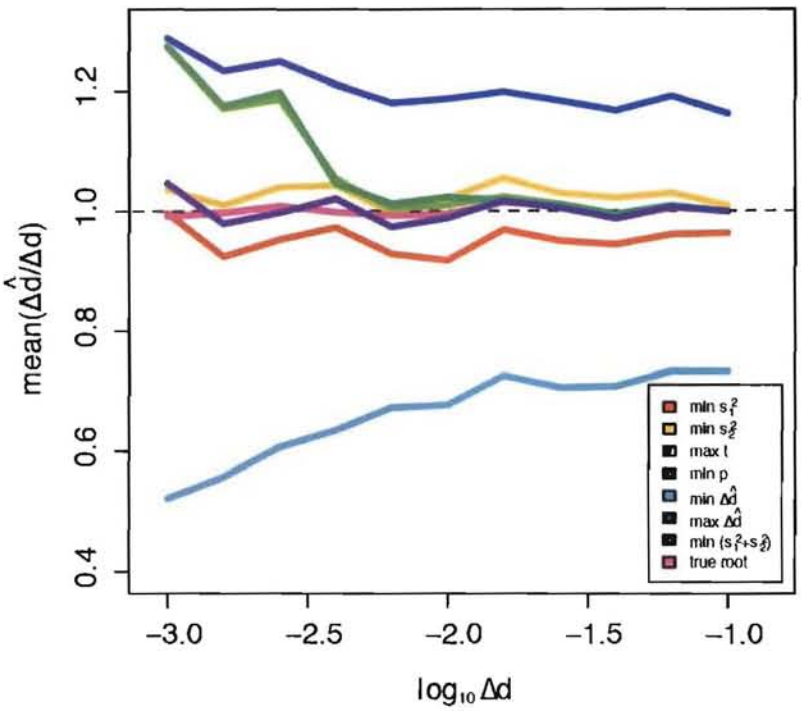


Figure S1. Comparison of different test statistics estimating the evolutionary rate. The dashed line indicates perfect estimation of $\Delta\hat{d}$ and the true root line shows the average estimated evolutionary rate at the true root, i.e., the theoretical limit of our simulations. All lines are simulations at $\Delta d/H=0.5$, and each data point is the average of 100 random simulated trees with 20 OTUs in each sample. Note that while the t and p value optimizations appear to be similar on average, in individual trees the best t value is not always the best p value because of how the degrees of freedom are calculated in Welch's t test (see Appendix A).

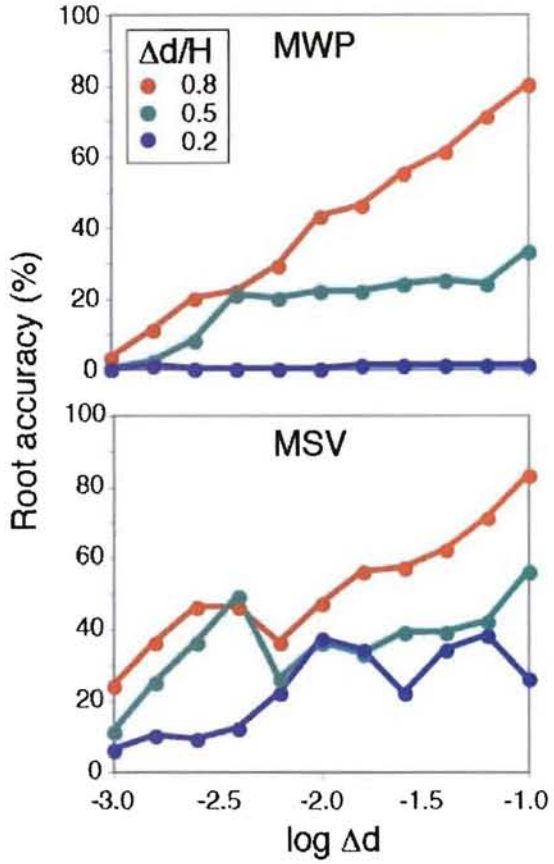


Figure S2. The ability to find the true root. Comparison of minimizing the sum of tip height variances of sample 1 and 2 (MSV) and minimizing Welch's p-value (MWP) as methods to find the true root. The different lines indicate different Δd/H ratios. Each data point is calculated as the number of times the true root was found out of 100 random simulated trees with 20 OTUs in each sample. Each tree was simulated with a defined root, then treated as unrooted, and run through our root and rate optimization to estimate the rooting point.

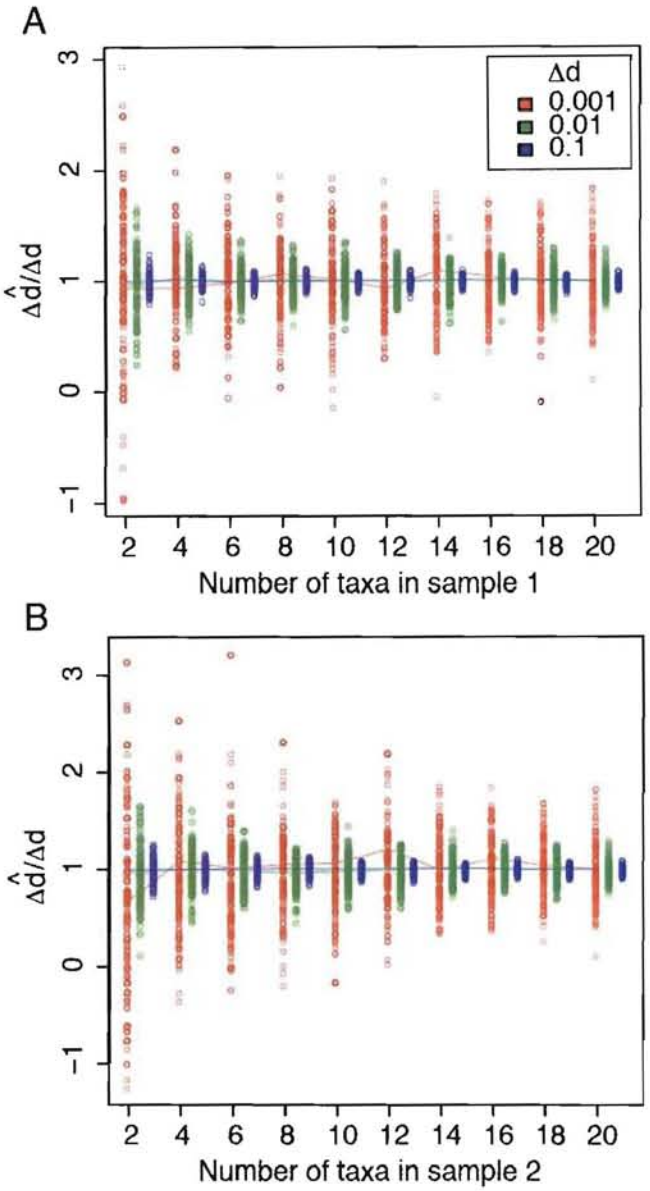


Figure S3. The effect of number of OTUs in sample 1 and 2. The colored lines show the average $\Delta\hat{d}$ estimates of the MSV optimality criterion, and open circles show the results from individual random trees (100 at each rate and OTU ratio). Simulations were done at three different expected rates (Δd=0.001, 0.01 and 0.1 substitutions/site), and at Δd/H=0.5. In panel **A** the number of OTUs in sample 1 varied according to the x-axis while the number was constant in sample 2 (N=20), and vise versa in panel **B**.

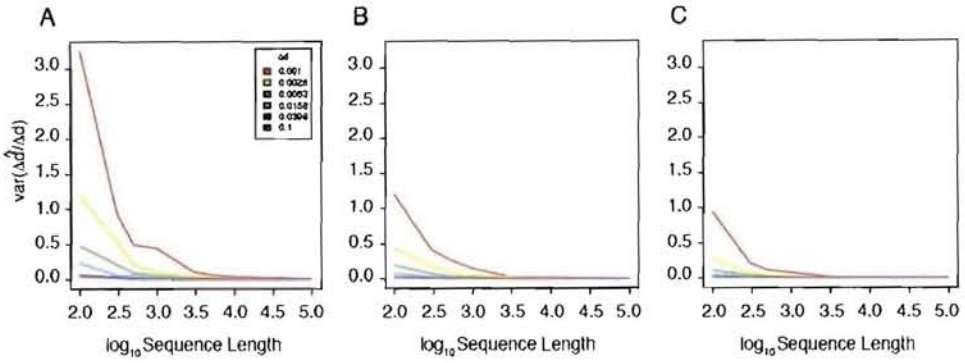


Figure S4. The effect of sequence length on the estimated rate $\Delta\hat{d}$. Since there was no bias in the MSV method, we show only the variance of the rate estimates as a function of sequence length. The different lines show the mean variance from 100 simulations in each data point at 6 different expected rates Δd, see inset box for color coding. Trees with 20 OTUs in each sample were used. Panel **A** is at Δd/H=0.2, **B** Δd/H=0.5, and **C** Δd/H=0.8.

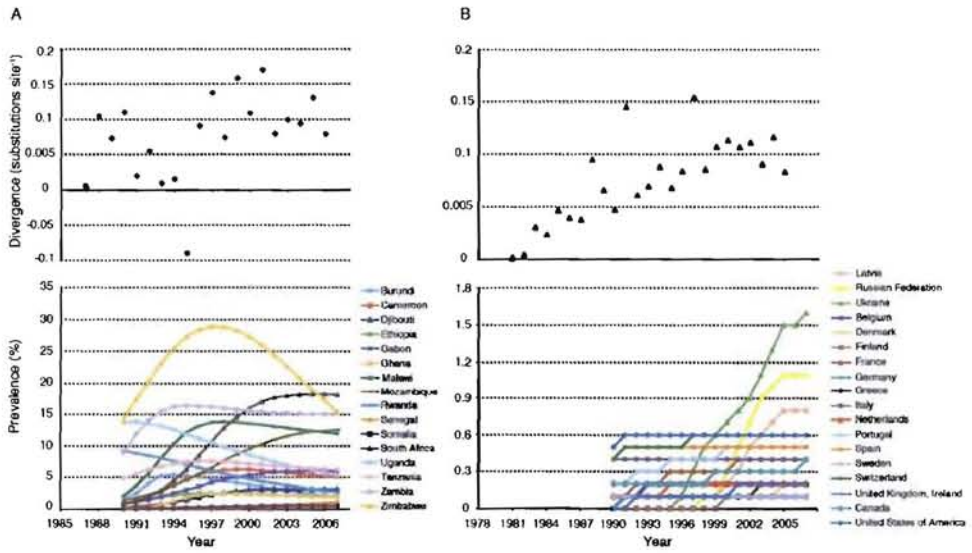


Figure S5. Tracking the dynamics of HIV-1 subtype B and C epidemics.
The temporal trend of the evolutionary rate on the population level (genetic divergence) of HIV-1 subtype C in Africa (**A**) was more variable than the temporal trend of HIV-1 subtype B in Europe and North America (**B**). Each divergence data point indicates the evolutionary rate calculated from a separate tree optimized by TreeRate. The divergence in both epidemics was calculated from the earliest available sequence samples, 1978+1979 for subtype B and 1984 for subtype C. HIV-1 prevalence data for countries

included in this study was derived from UNAIDS. Observe that the prevalence reflects all subtypes in the respective country, and may or may not be representative for subtypes B and C.