Dual-rate decomposition of phylogenetic lineages reveals epidemic dynamics of episodic fast and slow spread of HIV-1

Ethan Obie Romero-Severson[1,*], Nick Hengartner[1], Irina Maljkovic Berry[3], Namrata K. Patel[1], Edward Ionides[2], Thomas Leitner[1]

1 Los Alamos National Laboratory, Theoretical Biology and Biophysics

2 University of Michigan, Department of Statistics

3 Division of Infectious Diseases, The Feinberg School of Medicine, Northwestern University

∗ **E-mail: Corresponding eoromero@lanl.gov**

Abstract

The population-level evolutionary rate of HIV contains vital information about the rate of the epidemic spread due to the fact that the time between transmissions affects the population level divergence. A sequence of rapid transmissions between hosts correlates with a slow evolutionary rate, while a sequence of less frequent transmissions between hosts correlates with a fast evolutionary rate. In real epidemics these two modes of rapid and slow spread will be mixed across lineages in a corresponding phylogenetic tree as well as over time. In this paper we develop the 'dual-rate model' that allows us to identify regions in a phylogenetic tree that correspond to periods of fast and slow epidemic spread. We investigated two separate HIV-1 epidemics with very different spread patterns, the Latvian HIV-1 subtype A1 IDU and heterosexual epidemic and part of the Swedish HIV-1 subtype B IDU epidemic. First, using a Bayesian data augmentation model we show that even on level of whole lineages (root to tip) more than one evolutionary rate was present in the Latvian data, but not in the Swedish data. Next, we developed a two-state Markov chain model that assigns evolutionary rates along branches of a phylogeny by switching between *Slow* and *Fast* epidemiological states. We show that this dual-rate model accurately recovered the heterogeneity of spread rates across lineages and time in the Latvian A1 epidemic as well as recovering the more homogeneous and stable spread of HIV-1 amongst Swedish IDUs. Our method is able to reveal sub-branch *Fast* and *Slow* spread, related to individual chains of similar host-host spread rates. Thus, in this paper we show how the evolutionary rate – epidemic spread connection can be used to infer detailed and localized epidemic spread patterns that otherwise might have been missed in a phylogeny.

Author Summary

Because pathogen phylogenies are objective records of the pathogen epidemiology they have become a popular tool in reconstructing the spread of many measurably evolving pathogens, in particular many RNA viruses. We have previously shown that HIV-1 epidemics that have spread at different speeds display an inverse relationship with the population-level evolutionary rate of HIV-1, i.e., fast spread is associated with a slow evolutionary rate. Here we show that it is possible to infer dynamic switching from fast to slow and slow to fast spread throughout a phylogeny at the sub-branch level. The sub-branch level is important because we seldom have a complete sample from an epidemic, i.e., many phylogenetic branches contain several transmissions. Our the dual-rate model was able to accurately reconstruct dynamic and heterogeneous spread patterns in the Latvian HIV-1 epidemic as well as the much more homogeneous spread pattern of HIV-1 in Sweden. This type of inference expands the ability to infer how HIV, and other pathogens with similar evolutionary systems, spread in host populations.

Introduction

The relationship between genetic distance and calendar time is not straightforward. Nevertheless, methods have been developed to incorporate a range of assumptions about both the evolutionary and demogrpahic processes that determine the complex mapping of genetic distance to calendar time [1]. These methods have been sucessfully used to address important questions in the epidemiology of HIV including estimating the time of the origin of pandemic HIV [2–4], estimating the distribution of serial intervals [5, 6], and parameter estimation [7,8]. Most relaxed-clock phylogenetic methods treat the distribution of evolutionary rates over a pylogenetic tree as phenomenological property determined by the data. This assumption allows those methods to incorporate a very wide range of potential evolutionary rates on a tree with a small number of parameters, which is a decided benefit when the typical goal of such analyses is to estimate the time at which internal nodes are likely to have occurred. However, the variation in evolutionary rates is caused, in part, by properties of the underlying infectious dynamics of HIV in the sampled population that are fundamentally interesting to an epidemiologist. Treating the evolutionary rates as a nuisance to estimate divergence times occludes any attempt to make inference to the dynamics that generated the variance in evolutionary rates themselves. In this paper we propose a method to 'decompose' estimates of evolutionary rates in a phylogenetic tree into distinct, epidemiologically meaningful states and use this method to study the epidemiological properties of the underlaying transmission dynamics.

Within-host genetic diversification of the the V3 region of the HIV envelope occurs in distinct phases [9,10]. Initially mutations accumulate slowly limited by neutral evolution and possible reversions of mutations such as CTL escape mutations [11]; however, after the infected person mounts a robust response to the infection, the immune system drives evolution of the population of HIV-1 through a process of immunological escapes leading to an elevated evolutionary rate [10,12]. We refer to these two within-host periods as stage one and two respectively. Sequential transmissions where the time between transmissions is less than the length of stage one will show up as distended periods of particularly slow evolution as transmission occurs before the immune system has a chance to drive evolution of the viral population. Conversely, a sequence of transmissions where the time between transmissions is longer than stage one will show up as a period of faster evolution as the viral population in each infected host has been driven by diversifying selection. Similarly, for transmissions occurring later in the course of infection, the infected person will have passed into stage 2 and transmit a virus that has accumulated mutations at a faster rate.

Thus, the phylogenetic signal of evolutionary rates is correlated to the underlying rate of epidemic spread. This theoretical phenomenon has been directly observed in studies of epidemics believed to be dominated by either fast or slow epidemic spread [10] as well as in epidemics with overall changing spread rates [13]. If periods of slow and fast evolution can be identified on a phylogenetic tree, then we may be able to identify times and places where HIV-1 outbreaks happened or are presently occurring.

Materials and Methods

Sequence data

We used two sets of sequence data, one representing a mixed IDU/heterosexual and rapidly expanding outbreak in Latvia, and the other representing an ongoing IDU transmission chain in Sweden during a long period of stable incidence. For the Latvian data, HIV-1 DNA sequence data were collected and sampled as previously described [14–16]. We analyzed 271 HIV-1 subtype A1 sequences from the env V3 region with known sampling dates. The aligned and trimmed sequences had length 362 nt. The HIV-1 phylogeny was inferred using PhyML 3.0 [17] with the GTR + Γ + I evolutionary model and the subtree pruning and re-grafting search method. The tree was rooted at the MRCA of the set of samples from 1996-1998 (n=8) (figure 2). The coalescent times of the internal nodes were inferred using BEAST 1.7.5 [18] with the GTR + Γ + I evolutionary model assuming a fixed ML tree topology (from the PhyML reconstruction),

stepwise skyline population size model with 20 groups, and log-normal uncorrelated relaxed clock model. The Markov chain was run for 5×10^7 steps with samples taken every 2000 steps. A uniform prior was placed on the root height with bounds at 1982 and 1990 based on the first observation of an A1 infected person in Latvia in 1990. Branch lengths were estimated by taking the mean of each branch length from the posterior sample of trees with 20% of samples ignored as burn-in.

To isolate a Swedish transmission cluster that represents recent and historical transmission in Sweden we identified the largest clade of known Swedish HIV-1 subtype B IDU sequences: First, we found all the known HIV-1 subtype B envelope sequences with a known IDU risk factor that were isolated from a Swedish patient (http://www.hiv.lanl.gov/). Second, we used HIV BLAST (http://www.hiv.lanl.gov/) to find the 5 closet sequences, regardless of their geographic origin, to each index sequence. Third, to determine rough genetic relationships we made a neighbor joining tree of the set of unique sequences including all of the index patients and their closest BLAST results. We defined as the Swedish transmission cluster the largest subtree that contained only tips isolated from Sweden. Finally, we obtained the maximum likelihood topology and the genetic distances using PhyML 3.0 [17] under the GTR + Γ + I model (figure 3). The ML tree was rooted with the oldest sample from 1995, which was taken 5 years before the next oldest sample, as an outgroup. The times of the internal nodes was determined using the same method and assumptions as the Latvian data.

Sampling from the posterior of the data augmentation analysis

We take a Bayesian approach to estimate the model parameters by producing draws from the conditional distribution of the parameters and state variables given observed phylogenetic distance x_i^* and times y_i^* assuming the following prior distributions on the parameters

$$\begin{pmatrix} \mu_{Fast} \\ \mu_{Slow} \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Sigma}), \quad \kappa = \frac{1}{\tau} \sim \text{Gamma}(r, \lambda), \text{ and } \quad p \sim \text{Beta}(a_1, a_2)$$

An advantage of having introduced the unobserved state variables, is that we can readily obtain these draws using relatively simple Gibbs samplers. Specifically, we sequentially draw realizations from the conditional distributions: The state variables given data and values of the other parameter

$$\xi_i | x_i^*, y_i^*, \mu_{Fast}, \mu_{Slow}, \kappa, p \sim \text{Bernoulli}(Q_i),$$

with

$$Q_i = \frac{A_i}{A_i + B_i}$$

and

$$\begin{aligned} A_i &= p \times \exp\left(-\frac{\kappa}{2}(x_i^* + \mu_{Slow} \cdot y_i^*)^2\right) \\ B_i &= (1-p) \times \exp\left(-\frac{\kappa}{2}(x_i^* - \mu_{Fast} \cdot y_i^*)^2\right), \end{aligned}$$

the proportion p of lineages in the *Slow* state, given the augmented data and other parameters

$$p | \mathcal{A}, \mu, \kappa \sim \text{Beta}\left(a_1 + \sum_n \xi_i, a_2 + \sum_i (1 - \xi_i)\right),$$

the conditional distribution of the precision $\kappa = 1/\tau$, conditionally on the augmented data $\mathcal{A}$ and the parameters μ_{Fast} and μ_{Slow} is

$$\kappa | \mathcal{A}, \mu \sim \Gamma\left(r + \frac{n}{2},\ \lambda + \frac{1}{2}\sum_{i=1}^{n} \frac{(x_i^* - (\mu_{Fast} + (\mu_{Slow} - \mu_{Fast})\xi_i) \cdot y_i^*)^2}{y_i^*}\right).$$

To apply this model to the Latvian outbreak data, we set the hyper-parameters to the following values

$$\nu = \begin{pmatrix} 2.02 \times 10^{-3} \\ 16.9 \times 10^{-3} \end{pmatrix},\ \Lambda = \begin{pmatrix} 10^{-1} & 0 \\ 0 & 10^{-1} \end{pmatrix},\ r = 1,\ \lambda = 10^{-1},\ a_1 = a_2 = 1.$$

Constraints on the distribution of the evolutionary rates in the *Fast* state

To constrain the model we fixed μ_{Fast} and σ_{Fast}. Our constraints are based on an analysis of the evolutionary rates in the subtype A1 epidemic in IDUs in the former Soviet Republics (FSU), believed to be dominated by very fast transmission. Although estimates from this study are population based rather than on the within-host evolutionary rates, it stands to reason that in a very rapidly spreading epidemic in a susceptible population the time between transmissions will be low and, therefore, the population rate of evolution will be a reasonable prior on the evolutionary rate in the *Fast* state. The evolutionary rate in FSU was measured to be 0.20% substitutions/site with normally distributed error having standard deviation 0.02%. The point estimate was assumed to be a reasonable estimate of the mean prior distribution of evolutionary rates in the *Fast* state ($\mu_{Fast} = 0.2\%$. We selected 0.03% as the cut off point between the *Fast* and *Slow* states (at this value ¿99% of the density is in the interval $(0, 0.3\%]$). To lessen the effect of the low prior we assumed $\sigma_{Fast} = 0.2\%$ rather than 0.02%.

Iterated filtering

To determine reasonable values of θ we used the iterated filtering method `mif` in the R library `pomp 0.43-8` [**?**] to find an estimate of the maximum likelihood parameter set, θ_{MLE}. We treated the tree as a pseudo-timeseries where branches (organized in a depth-first fashion) constituted 'times' and the data were the external estimates of the branch lengths in time. For each branch the likelihood was calculated (algorithm 1) for a θ drawn from the filtering distribution. The method was run using 300 particles filtered over 300 iterations with a cooling rate of 0.985. We assumed μ_{Fast} and σ_{Fast} were fixed at values 0.02%. The method was run separately for the Latvian and Swedish data from several random starting points to avoid local optimization.

Tree paining

We refer to the distribution of *Fast* (red) and *Slow* states (blue) and their corresponding evolutionary rates over the tree to be the tree 'paining'. To paint the trees, first θ_{MLE} was found using an iterating filtering method. Then, for each branch in a depth-first order, the dual-rate model was realized 10,000 times with the parameters given by θ_{MLE}. The most likely realization of the model was assumed to be the paining for that branch. The evolutionary rate and state at the end of the parent branch was assumed to be the starting evolutionary rate and state at the beginning of the child branches. This process was repeated 100 times to generate 100 independent tree paintings each for the Latvian and Swedish data. The standard deviation of the likelihoods of the 100 paintings was 0.6 log units for the Latvian data and ¡0.1 log units for the Swedish data.

Results

Conceptual model and nomenclature

All of the analyses that we present in this paper are based on a simple conceptual 2-state model where any point along each branch in a phylogeny is in either the *Fast* state corresponding to rapid transmission between individuals and corresponding low rate of evolution or in the *Slow* state corresponding to slow transmissions between individuals and a much higher rate of evolution. Past work [Reference] suggests

we can constrain that evolutionary rates by $\lambda_{Slow} \leq c \leq \lambda_{Fast}$, for some known cut-off value of c. Our analysis will further assume the evolutionary rates on each branch to be random with with densities f_{Slow} and f_{Fast}, respectively. Specifically, we will assume that f_{Slow} is a Gaussian density $f(\lambda|\mu_{Slow}, \sigma_{Slow})$ with mean μ_{Slow} and standard deviation σ_{Slow} restricted to the internal $(0, c)$, and f_{Fast} is a Gaussian density $f(\lambda|\mu_{Fast}, \sigma_{Fast})$ restricted to the the half-line (c, ∞). In practice, the values of $\mu_{Slow}, \mu_{Fast}, \sigma_{Slow}$ and σ_{Fast} are such that the probability of being outside the restricted sets is essentially zero.

The data are phylogenetic tress with branch lengths measured in either genetic or temporal distances. Given a phylogenetic tree $\mathcal{T}$, with tips $\{1 \cdots n\}$ sampled at times $\{t_1, \ldots, t_n\}$ and internal nodes $\{n+1, \ldots, 2n-1\}$, and rooted at node $n+1$, assume that the distance from t_{n+1} to t_i is x_i^* and y_i^* measured in genetic and temporal distance respectively. $\mathcal{T}$ has edges $\{1, \ldots, 2n-2\}$ such that the length of the i^{th} edge has genetic length x_i and temporal length y_i.

Evidence for dual evolutionary rates in the Latvian and Swedish epidemics

Our first analysis looks for evidence of multiple evolutionary rates at the level of whole lineages (path from the root to the tip of each taxa) in a phylogenetic tree of the Latvian sequence data. To do this, we constructed a Bayesian data augmentation model that assumes the existence of an unobserved state variable, ξ_i, which for each extant taxa that takes the value 0 if the lineage is in the *Fast* state or 1 if the lineage is in the *Slow* state. In this framework, the (full) augmented data are

$$\mathcal{A} = \{(x_i^*, \xi_i, y_i^*), i = 1, \ldots, n\},$$

corresponding to the length of the lineage measured from root to tip in both genetic distance, x_i^*, and time, y_i^*, and the unobserved state variable ξ_i.

We assume that the conditional distribution of the phylogenetic distance x_i^* given time y_i^* and state variable ξ_i is Gaussian with conditional expectation expectation

$$\begin{aligned} \mathbb{E}[x_i^*|y_i^*, \xi_i] &= (\mu_{Fast} + (\mu_{Slow} - \mu_{Fast})\xi_i) \cdot y_i^* \\ &= \begin{cases} \mu_{Fast} \cdot y_i^* & \text{if } \xi_i = 0 \\ \mu_{Slow} \cdot y_i^* & \text{if } \xi_i = 1 \end{cases} \end{aligned}$$

and conditional variance variance

$$\text{Var}(x_i^*|y_i^*, \xi_i) = \tau y_i^*.$$

Given the high evolutionary rate in the V3 region that we used to infer the tree and the relatively long duration of the outbreak (¿10 years), the normal distribution is a reasonable approximation to the underling discrete process of accumulating mutations over time. We model the marginal distribution of the state by setting $\mathbb{P}[\xi_i = 1] = p$, that is, we assume that a fraction p of the rates are in the *Slow* state.

The prior mean evolutionary rates, ν, was informed by previously described average evolutionary rates in rapid spread among IDUs in former Soviet union, and slower spread among heterosexuals in Africa [10]. We used MCMC with Gibbs sampling to integrate over the posterior to get an estimate of μ_{Fast} and μ_{Slow} evolutionary rates and the probability that each lineage is in the *Fast* state. Figure 1 shows estimates of the evolutionary rates and the probability of assignment of each lineage to the high evolutionary rate group. In the Latvian epidemic both $\mu_{Fast} = 0.68\%$ and $\mu_{Slow} = 1.27\%$ substitutions/site are closer to one another than they are to the mean prior rates. This homogenizing of the evolutionary rates is expected given the fact that the lineages are not independent (branches near the root are shared by multiple lineages) and that a given lineage could represent a very large number of transmissions, possibly spanning years, that could contain both periods of fast and slow spread (e.g. a transmission into a new susceptible social network). However, even at this course-grained level of analysis there is evidence that two rates distinct evolutionary rates in the Latvian epidemic.

The estimated distribution of states in the Swedish epidemic is nearly all *Fast* with rates $\mu_{Fast} = 0.51\%$ and $\mu_{Slow} = 1.70\%$ substitutions/site. The high value μ_{Fast} in the Swedish data is effecting some degree

of overdispersion in the otherwise Poisson fit to the data (Figure 1). Thus, in this case there is little evidence for two highly distinct evolutionary rates.

The number of diagnoses in Swedish IDU epidemic is stable over the time that is covered by the Swedish tree while the infection rates in the Latvian are rapidly changing over the course of the time covered by the Latvian tree (Fig 4). This analysis shows that the distribution of evolutionary rates, even at the level of whole lineages, is influenced by the underlying epidemiology. To relax the strongest assumptions (independently evolving lineages and lineages as the basic analytical unit) we developed a dynamic and sub-branch dual-rate model.

The dual-rate model of HIV evolutionary rates

The data augmentation analysis in the previous section gives some evidence for more than one evolutionary rate in some contexts; however, transitions between fast and slow rates of epidemic spread can occur at the sub-branch level. For example, an infected person entering a highly connected IDU or sexual network could spark a small outbreak leading to a sequence of rapid transmissions that would show up as a period of fast spread at the sub-branch or branch level. Therefore, each branch in a tree should be able to have a unique pattern of transitions between *Fast* and *Slow* states. Further, because the evolutionary rates in periods of *Fast* and *Slow* states are determined by the virus-host interaction over one or more transmissions, the variance of evolutionary rates in each state should have a non-zero value to account for unobserved virus and host heterogeneity. To implement this model we used a simple two-state Markov chain that assigns evolutionary rates along branches of a phylogeny by switching between the *Slow* and *Fast* states at given rates (Fig 9). The vector $\theta = \{\mu_{Slow}, \sigma_{Slow}, \mu_{Fast}, \sigma_{Fast}, \delta_{S\to F}, \delta_{F\to S}\}$ specifies the model where $\delta_{S\to F}$ is the rate of switching from the *Slow* to *Fast* state, and $\delta_{F\to S}$ is the rate of switching from the *Fast* to *Slow* state. The model assumes that each edge 'inherits' both the initial state at the end of its parent edge and the evolutionary rate, λ_i, such that the initial condition of the model is only the state at the root of the phylogenetic tree. To constrain the model to a more reasonable number of parameters, we fix μ_{Fast} and σ_{Fast} to a universal within-patient level (detailed in Methods section).

The data for the dual-rate model is a set of branch lengths of a phylogenetic tree with a fixed topology measured in both units of calendar time and genetic distance. In general, genetic branch lengths are inferred using a standard evolutionary model and temporal branch lengths are either directly observed, such as through a known introduction event or transmission, or inferred through relaxed-clock phylogenetic methods [1,19] The dual-rate model translates the genetic distance of each branch into a temporal distance conditional on θ by first selecting a value of θ and then simulating the dual-rate model over the length of the branch. The simulation produces a step function over the branch alternating between evolutionary rates drawn from $f(\lambda|\mu_{Slow}, \sigma_{Slow})$ and $f(\lambda|\mu_{Fast}, \sigma_{Fast})$. The reciprocal of λ is the rate at which time accumulates as a function of genetic distance, therefore integrating over the reciprocal of the simulated step function gives the length of the branch in calendar time. The observed branch length in calendar time is y_i and the simulated branch length is $\hat{y_i}$, therefore the likelihood of branch i is $\mathcal{L}(y_i|\theta) \sim \mathcal{N}(y_i|\mu = \hat{y_i}, \sigma = 1)$. The simulation algorithm is shown in figure 1.

Dual rate tree patterns reveal different spread dynamics in Latvian and Swedish IDU networks

The maximum likelihood 'paintings' (i.e. the distribution of *Fast* and *Slow* states over a time-scaled phylogeny depicted as colors) of the Latvian and Swedish data are shown in figures 5 and 7. In the single best painting of the Latvian phylogenetic tree we found the average evolutionary rate to be 0.09% in the *Fast* and 2.07% substitutions/site in the *Slow* state. Likewise, in the Swedish epidemic we found the average evolutionary rate to be 0.1% in the *Fast* and 1.75% substitutions/site in the *Slow* state.

The Latvian tree shows clearly clustered regions of mostly *Fast* or *Slow* spread suggesting that the epidemic is quite heterogeneous with respect to the rate of spread and that lineages of fast spread can exist

for several years. The heterogeneity of the Latvian tree painting is consistent with the epidemiology of HIV in Latvia [14–16]. Figure 4 shows the number of HIV diagnoses stratified by risk factor (heterosexual, IDU). The very rapid increase in IDU cases must have been driven by rapid sequential, and parallel, transmissions, as indicated by the high predominance of blue (*Fast*) near the root of the tree. Heterosexual transmission in Latvia also appears to be approximately exponentially increasing although at a much slower rate, which is represented by long red branches of het-het transmissions. This tree painting is consistent with the theory that the HIV-1 subtype A1 epidemic in Latvia rapidly spread through the IDU population and spilled over into the heterosexual population where it sustained by slower, but with sufficient force of infection to be above the epidemic threshold, het-het chains of transmission [16]. If this theory is correct, then focusing on prevention efforts on the higher-risk IDU population will not necessarily translate to prevented infections in the at-risk heterosexual population [20].

The Swedish tree painting gives a very different picture, however, also consistent with the dynamics among Swedish subtype B IDU infected [21]. The inferred evolutionary rates in the Swedish tree are much closer to one another than in the Latvian tree, which suggests a more homogeneous process of spread (i.e. on average, transmissions occurring in the *Slow* state are not that much slower than in the *Fast* state). Thus, this result is consistent with both the data augmentation analysis and the known epidemiology. The diagnosis rate of new IDU cases in Sweden over the time spanned by the tree is nearly constant excluding an slight increase in new diagnoses in 2006-2007 (figure 4. The low number of total cases and the constant diagnoses rate implies a mostly homogeneous epidemic without significant fast and slow spreading substructure like that observed in the Latvian A1 tree. However, the slight increase in diagnosed cases in 2006-2007 aligns with a cluster of rapid spread beginning in early 2003.

This method cannot directly estimate the proportion of transmissions that are caused by recently infected persons, but the average proportion of time spent in the *Fast* state on extant lineages can be thought of as a lower bound on this key epidemiologic parameter. Figures 6 and 8 show the the average proportion of time spent in the *Fast* state on extant lineages over time in the Latvian and Swedish trees respectively. The Latvian tree shows a very high proportion of *Fast* very early on in the epidemic dropping down to about 50-60% around 2005. The first cases of HIV were diagnosed in Latvia in the mid 1990s and rapidly spread through the IDU population reaching a zenith in 2001 only about 4 years after the the first known IDU case. The extremely rapid early spread among IDUs must correspond to a very high proportion of transmissions from recently infected people, which the results from the dual-rate model clearly shows. Likewise, as the force of infection rapidly drops off and the epidemic stabilizes our dual-rate inference indicates a corresponding stabilization in the proportion of lineages in the *Fast* state, suggesting that the tree painting is consistent with theoretical expectations given the epidemiology of HIV in Latvia [16]. The Swedish epidemic is stable over the course of time covered by the tree, and the dual-rate inference reports a corresponding flat proportion of of time spent in the *Fast* state.

Discussion

We have previously shown that the HIV molecular clock rate in a phylogeny is inversely correlated to the HIV spread rate among its hosts [10, 13]. The overall evolutionary rate at any time across the entire tree may, however, mask individual deviations from the general trend, such as relatively small outbreaks or heterogeneous spread patterns in parallel transmission chains associated with different risk groups. From a public health and prevention perspective, such mixed patterns may in fact be the most important epidemiological information one wants to extract. In this paper we have shown that the pattern of how the molecular clock rates are distributed across a phylogeny is indeed revealing how HIV spreads in a host population. In the Latvian population we accurately reconstructed the early fast spread among IDU of HIV-1 subtype A1, and the later, and generally slower, spillover and spread among heterosexual contacts. Our dual-rate inference method also accurately reconstructed the quite different spread pattern among Swedish HIV-1 subtype B infected IDUs.

The difference between *Fast* and *Slow* spread is not fixed or absolute, as seen in the inference of the Swedish IDU data. Ideally, *Fast* spread is only involving transmissions during the early phase of infection, when the immune pressure has not yet strongly affected the within-host evolutionary rate, and *Slow* spread only involves spread during the later infection phase. In real epidemics we will likely see mixtures of these two idealized rates. Thus, *Slow* spread in one case may in addition to transmissions during later stage infections also involve a few transmissions shortly after infection, while in another case there may be more fast transmissions affecting the *Slow* rate. Similarly, *Fast* spread may involve different ratios of fast and slow spread. Thus, *Slow* and *Fast* in two different situations may mean different things. In fact, comparing the posterior rates to the priors (which were estimated from fairly homogeneous populations [10]) and to each other gives additional information about the heterogeneity of spread rates in the population studied.

Our method operates on the sub-branch level, which to some extent takes into account that we have a sparse sample from the epidemic. This means that many individual hosts may be represented by a single phylogenetic branch [22]. While it must be true that any transition from red (*Slow*) to blue (*Fast*) in the 'painted' trees involve one host-host transmission, the reverse (blue to red) does not neccessarily mean one transmission, as a host itself transitions from early to late phase infection. Furthermore, sustained *Fast* transmissions (as in the early Lativan A1 epidemic, Fig 5) will cause relatively long blue branches or branch segments – such streaks obviously involve many hosts. Similarly, red branch segments may involve several later phase (chronic) transmissions. Thus, blue and red segments generally are likely to contain ¿1 host of similar stage transmissions. It is also important to point out that while a slow evolutionary rate implies spread during the initial infection phase, it does not necessarily imply that massive numbers of hosts got infected. The Swedish IDU epidemic we examine here shows that the HIV-1 population-level evolutionary rate is slow, indicating that most new IDU infections came from recently infected donors, while the incidence is still quite low. This can be explained by a small number of active transmission chains (e.g. non-terminating chains). Thus, a slow evolutionary rate indicates that transmissions occur in the acute disease stage, and if not many people become infected then there is potential for it. In the case of IDU mediated spread, this pattern means that treatment as prevention and sero-sorting will not be effective to stop new infections, but needle exchange programs potentially would.

Because the spread rate directly influences the population level evolutionary rate [10], one must be careful not to make assumptions about the evolutionary process that would affect the epidemiological interpretation of the phylogenetic results from an epidemic. Temporal data could come from independent information such as recorded outbreak starts or when someone infected another, or data about when patients were infected based on independent biomarker data [23]. When such data is unavailable, or only partially available, one can infer transmission times using molecular clock approaches. When doing that it is crucial to avoid biasing the molecular clock in such a way that it reduces local deviations or homogenizes rates across branches. Relaxed or local clocks would be proper for this purpose.

As the marriage of phylogenetics and dynamic epidemiological modeling matures, phylodynamic studies will become more powerful and able to extract epidemiological information that traditional epidemiological methods have difficulties with. Because HIV phylogenies are affected by host-host transmissions, there has been quite a lot of interest to use HIV genetic information. However, the connection between the evolutionary rate of the ethiological agent and the spread rate among hosts is one source of information that as of yet has not been extensively used [10, 13, 24]. In this paper we explored further how this connection can be used to infer heterogeneous epidemic patterns that otherwise might have been missed if one would have averaged across the phylogeny. Future developments may include within-host dynamics and between host transmissions to accurately model effective clock rates and also include the effects of the pre- transmission interval [12, 22].

Acknowledgments

References

References

1. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS biology 4: e88.

2. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, et al. (2000) Timing the ancestor of the HIV-1 pandemic strains. Science (New York, NY) 288: 1789–1796.

3. Lemey P, Pybus OG, Rambaut A, Drummond AJ, Robertson DL, et al. (2004) The molecular population genetics of HIV-1 group o. Genetics 167: 1059–1068.

4. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, et al. (2008) Direct evidence of extensive diversity of HIV-1 in kinshasa by 1960. Nature 455: 661–664.

5. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Brown AJL (2008) Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS medicine 5: e50.

6. Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, et al. (2009) Molecular phylodynamics of the heterosexual HIV epidemic in the united kingdom. PLoS Pathogens 5: e1000590.

7. Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, et al. (2011) Transmission network parameters estimated from HIV sequences for a nationwide epidemic. Journal of Infectious Diseases 204: 1463–1469.

8. Volz EM, Ionides E, Romero-Severson EO, Brandt MG, Mokotoff E, et al. (2013) HIV-1 transmission during early infection in men who have sex with men: A phylodynamic analysis. PLoS Med 10: e1001568.

9. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, et al. (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. Journal of virology 73: 10489–10502.

10. Maljkovic Berry I, Ribeiro R, Kothari M, Athreya G, Daniels M, et al. (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. Journal of virology 81: 10625–10635.

11. Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, et al. (2004) HIV evolution: CTL escape mutation and reversion after transmission. Nature Medicine 10: 282–289.

12. Leitner T, Albert J (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proceedings of the National Academy of Sciences of the United States of America 96: 10752–10757.

13. Maljkovic Berry I, Athreya G, Kothari M, Daniels M, Bruno WJ, et al. (2009) The evolutionary rate dynamically tracks changes in HIV-1 epidemics: application of a simple method for optimizing the evolutionary rate in phylogenetic trees with longitudinal data. Epidemics 1: 230–239.

14. Balode D, Ferdats A, Dievberna I, Viksna L, Rozentale B, et al. (2004) Rapid epidemic spread of HIV type 1 subtype a1 among intravenous drug users in latvia and slower spread of subtype b among other risk groups. AIDS research and human retroviruses 20: 245–249.

15. Balode D, Skar H, Mild M, Kolupajeva T, Ferdats A, et al. (2012) Phylogenetic analysis of the latvian HIV-1 epidemic. AIDS research and human retroviruses 28: 928–932.

16. Graw F, Leitner T, Ribeiro RM (2012) Agent-based and phylogenetic analyses reveal how HIV-1 moves between risk groups: injecting drug users sustain the heterosexual epidemic in latvia. Epidemics 4: 104–116.

17. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology 59: 307–321.

18. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology 7: 214.

19. Drummond AJ, Suchard MA (2010) Bayesian random local clocks, or one rate to rule them all. BMC biology 8: 114.

20. Koopman JS, Simon CP, Riolo CP (2005) When to control endemic infections by focusing on high-risk groups. Epidemiology (Cambridge, Mass) 16: 621–627.

21. Skar H, Axelsson M, Berggren I, Thalme A, Gyllensten K, et al. (2010) Dynamics of two separate but linked HIV-1 CRF01_ae outbreaks among injection drug users in stockholm, sweden, and helsinki, finland. Journal of Virology 85: 510–518.

22. Romero-Severson E, Skar H, Bulla I, Albert J, Leitner T (2014) Timing and order of transmission events is not directly reflected in a pathogen phylogeny. Molecular Biology and Evolution : msu179.

23. Skar H, Albert J, Leitner T (2013) Towards estimation of HIV-1 date of infection: a time-continuous IgG-model shows that seroconversion does not occur at the midpoint between negative and positive tests. PloS one 8: e60906.

24. Duffy S, Shackelton LA, Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. Nature Reviews Genetics 9: 267–276.

Figure Legends

Figure 1. Bayesian data augmentation classification of lineages into *Fast* and *Slow* spreading states. This figure shows the average posterior assignment of each lineage to the *Fast* (blue) and *Slow* (red) states in the Latvian data (top) and Swedish data (bottom). In both cases the priors on the evolutionary rates in the *Fast* and *Slow* states were Normal with means 2.02% and 16.9% per site per year respectively.

Figure 2. Phylogeny of 271 V3 sequences from the mixed IDU-heterosexual Latvian epidemic. This is the maximum likelihood phylogeny of 271 V3 HIV-1 sequences sampled from 1998 through 2007 infered under the GTR + Γ + I model. The sample year of each tip is indicated by its color.

$\alpha \leftarrow 0$
$\hat{y} \leftarrow 0$
while $\alpha < x$ **do**
    if S is $Fast$ **then**
        draw χ from Exponential($\delta_{F \to S}$)
        if $\alpha + \chi < x$ **then**
            $\hat{y} \leftarrow \alpha + \chi \mu_{Fast}^{-1}$
            $\alpha \leftarrow \alpha + \chi$
        else
            $\hat{y} \leftarrow (x - \alpha) \mu_{Fast}^{-1}$
            $\alpha \leftarrow x$
        end if
        $S \leftarrow Slow$
    end if
    if S is $Slow$ **then**
        draw χ from Exponential($\delta_{S \to F}$)
        if $\alpha + \chi < x$ **then**
            $\hat{y} \leftarrow \alpha + \chi \mu_{Slow}^{-1}$
            $\alpha \leftarrow \alpha + \chi$
        else
            $\hat{y} \leftarrow (x - \alpha) \mu_{Slow}^{-1}$
            $\alpha \leftarrow x$
        end if
        $S \leftarrow Fast$
    end if
end while
$\mathcal{L}(e|\theta) \leftarrow \mathcal{N}(y_i | \mu = \hat{y}_i, \sigma = 1)$

Algorithm 1. Method for calculating the likelihood of a given edge, e, in a phylogenetic tree with lengths y in calendar time and x in genetic units given the dual-rate model, θ, and the starting condition, S.

Figure 3. Phylogeny of 39 V3 sequences from a Swedish IDU cluster. This is the maximum likelihood phylogeny of 39 V3 HIV-1 sequences sampled from 1998 through 2007 infered under the GTR + Γ + I model. The sample year of each tip is indicated by its color.

Figure 4. Number of diagnoses of HIV in Latvia and Sweden. The yearly number of diagnosed cases among Latvian IDUs (blue dashed), Latvian heterosexuals (red dot-dashed), and Swedish IDUs. The Latvian IDU and heterosexual diagnosed cases are almost exclusively subtype A1 while the Swedish IDUs are mostly subtype B.

Figure 5. A time-scaled phylogenetic tree of the Latvian epidemic 'painted' with periods of fast and slow epidemic spread. Blue represents time spent in the *Fast* state (fast epidemic spread and slow evolutionary rate) where transmissions are occurring within 6 months of infection, while red represents time spent in the *Slow* state (slow epidemic spread and fast evolutionary rate). Branch lengths are measured in calendar time and span the period 1997-2007.

Figure 6. Average proportion of extant lineages in the *Fast* state in the Latvian epidemic. This figure shows the average of all the proportions of extant lineages currently in the *Fast* state (fast epidemic spread, slow evolutionary rate). Grey lines show the average proportion from each of 50 optimizations, the red line shows the average proportion averaged over each of the 50 optimizations.

Figure 7. A time-scaled phylogenetic tree of the Swedish IDU cluster 'painted' with periods of fast and slow epidemic spread. Blue represents time spent in the *Fast* state (fast epidemic spread and slow evolutionary rate) where transmissions are occurring within 6 months of infection, while red represents time spent in the *Slow* state (slow epidemic spread and fast evolutionary rate). Branch lengths are measured in calendar time and span the period 1997-2007.

Figure 8. Average proportion of extant lineages in the *Fast* state in the Swedish cluster. This figure shows the average of all the proportions of extant lineages currently in the *Fast* state (fast epidemic spread, slow evolutionary rate). Grey lines show the average proportion from each of 50 optimizations, the red line shows the average proportion averaged over each of the 50 optimizations.

Tables

Supporting Information Legends

Figure 9. Illustration of the dual-rate HIV clock model. A) An illustration of the Markov chain used to represent the dual-rate HIV clock model. At any point in time a lineage is in either in the *Fast* state (short times between transmissions, slow evolution) or in the *Slow* state (longer times between transmission, fast evolution). The evolutionary rates in the *Fast* and *Slow* are drawn from $f(\mu_{Fast}, \sigma Fast)$ and $f(\mu_{Slow}, \sigma Slow)$ respectively and are restricted to not overlap, that is, the highest evolutionary rate in the *Slow* state cannot be higher than the lowest evolutionary rate in the *Fast* state. B) Illustrates an example of the dual-rate model over a possible transmission genealogy. C) Represents the implied genetic distances given the particular realization in (B).