Computational Inference Methods for Selective Sweeps Arising in Acute HIV Infection

Sivan Leviyang¹

Department of Mathematics and Statistics, Georgetown University, Washington, DC 20057

ABSTRACT During the first weeks of human immunodeficiency virus-1 (HIV-1) infection, cytotoxic T-lymphocytes (CTLs) select for multiple escape mutations in the infecting HIV population. In recent years, methods that use escape mutation data to estimate rates of HIV escape have been developed, thereby providing a quantitative framework for exploring HIV escape from CTL response. Current methods for escape-rate inference focus on a specific HIV mutant selected by a single CTL response. However, recent studies have shown that during the first weeks of infection, CTL responses occur at one to three epitopes and HIV escape occurs through complex mutation pathways. Consequently, HIV escape from CTL response forms a complex, selective sweep that is difficult to analyze. In this work, we develop a model of initial infection, based on the well-known standard model, that allows for a description of multi-epitope response and the complex mutation pathways of HIV escape. Under this model, we develop Bayesian and hypothesis-test inference methods that allow us to analyze and estimate HIV escape rates. The methods are applied to two HIV patient data sets, concretely demonstrating the utility of our approach.

ACUTE HIV-1 infection is marked by an initial period of 2–4 weeks in which the viral population expands from one to five infected cells to $\sim 10^9$ infected cells. Following this expansion period, in the subsequent 1–2 months, the viral load drops and reaches a setpoint (Fiebig *et al.* 2003; Mehndru *et al.* 2004, 2007; Keele *et al.* 2008). Cytotoxic T lymphocytes (CTLs) are thought to play an important role in shaping acute infection (Goulder and Watkins 2004; Cohen *et al.* 2011). The onset of CTL response is temporally correlated with the end of the expansion period, suggesting a role for CTLs in controlling viral load. Numerous studies have shown that during acute infection, specific HIV mutations on CTL targeted epitopes sweep to fixation, providing more direct evidence that CTL response shapes the infecting HIV population; see Goulder and Watkins (2004) for a review.

In recent years, full-genome sequencing studies have provided an increasingly detailed description of CTL response during acute infection, *e.g.*, Fernandez *et al.* (2005), Goonetilleke *et al.* (2009), Liu *et al.* (2006), Henn *et al.* (2012), Bimber *et al.*

(2009), and see Boutwell *et al.* (2010) for a review. A general picture in which initial CTL response targeting a single epitope occurs several days before peak viral load has emerged. Shortly after this initial CTL response, CTL responses at one to three other epitopes occur.

Recent deep-sequencing data sets have provided a picture of HIV escape at the CTL targeted epitopes, *e.g.*, Fisher *et al.* (2010), Henn *et al.* (2012), Bimber *et al.* (2009). HIV escape mutations at the first CTL targeted epitope rise to significant proportions 1–3 weeks after peak viral load. Escape mutations at the next series of epitopes targeted rise to significant proportion within roughly 4–6 weeks of peak viral load. Importantly, deep-sequencing studies have shown that HIV escape at a targeted epitope often occurs along multiple mutation pathways. The different mutation pathways are simply different nucleotide substitutions in the targeted epitope. During an HIV escape, these different mutations sweep to significant frequencies simultaneously, thereby replacing an HIV population typically homogeneous at the targeted epitope with a population possessing multiple, significant frequency, mutation variants at the epitope. This phenomenon, termed “epitope shattering” (Boutwell *et al.* 2010), occurs at each of the multiple epitopes targeted by CTLs, producing an enormous array of possible mutation pathways.

Quantifying HIV escape rates may help clarify the role of CTL response in controlling and shaping infection. However,

Copyright © 2013 by the Genetics Society of America

doi: 10.1534/genetics.113.150862

Manuscript received February 26, 2013; accepted for publication May 5, 2013

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.150862/-/DC1>.

¹Address for correspondence: Mathematics and Statistics Department, Georgetown University, St. Mary's Hall, 3rd Fl., Washington, DC 20057. E-mail: sr286@georgetown.edu

Figure 1 Single-escape graph for CH40.

epitope shattering, especially when considered across multiple targeted epitopes, makes modeling and quantitative analysis of HIV escape difficult. Techniques exist to infer HIV escape rates (Fernandez *et al.* 2005; Asquith *et al.* 2006; Ganusov and De Boer 2006; Goonetilleke *et al.* 2009; Ganusov *et al.* 2011) and these techniques have been valuable in quantifying the role of CTL killing in acute and chronic infection. However, these techniques consider a single epitope and assume only two HIV variants at that epitope, a wild type and mutant type. The effect of these assumptions on escape-rate estimates is not known.

Put in evolutionary terms, HIV escape from CTL response forms an example of a selective sweep. A range of inference methods exist to analyze selective sweeps (*e.g.*, Kaplan *et al.* 1988; Tajima 1989; Gillespie 1991; Krone and Neuhauser 1997; Nielsen and Yang 1998), but these inference methods assume models that do not fit the specific biology of acute HIV infection. Several factors make HIV evolution during acute infection nonstandard. First, the HIV population size is not fixed during initial infection (Stafford *et al.* 2000; Fiebig *et al.* 2003). Second, since CTL populations change over the time period of the sweep, the selective force exerted by CTLs is time varying. Third, the selective force exerted on the HIV population by CTL targeting and other fitness effects differs across the many variants involved in escape, producing a complex form of selection. Fourth, HIV escapes selection through multiple mutation pathways. While models that reflect some of these forces have been constructed (for example, see Slatkin and Hudson 1991 for a model of an exponentially growing population applicable to the HIV expansion time period), combining all these evolutionary effects in the context of inference has not been explored. Several authors have examined selection in the context of HIV (*e.g.*, Nielsen and Yang 1998; Frost *et al.* 2000; Pennings and Hermisson 2006; Rouzine and Coffin 2010; Batorsky *et al.* 2011), but not with models that reflect the unique features of CTL response and HIV escape during acute infection.

In this work, we describe a model and associated computational methods accounting for multi-epitope CTL response and varied HIV escape pathways. Our focus is on modeling and inferring escape rates. The model is based on the standard model of viral dynamics (Perelson 2002). However, since HIV mutants produced in epitope shattering do not exist at initial infection and often arise sequentially in time, we extend the deterministic standard model to a stochastic birth–death process that includes mutation. Intuitively, the birth–death process is an agent-based system that tracks the birth, death, and mutation events of individual infected cells. Such an approach

Figure 2 Double-escape graph for CH58.

to HIV dynamics has been used previously (*e.g.*, Ribeiro and Bonhoeffer 1999; Tuckwell *et al.* 2000; Zhuo and Dorman 2005; Merrill 2005; Tuckwell *et al.* 2008).

Accounting for the many HIV variants found in the deep-sequencing data sets mentioned can lead to high-dimensional models with many parameters, a setting for which inference is computationally difficult and often highly sensitive to the parameter values chosen. With this in mind, the focus of our inference methods is on hypothesis testing. We develop a computational method for testing whether a data set is in a sense typical of data produced from a particular model and set of parameter values. Rather than building confidence intervals that describe the most likely parameters that generated the data set, we ask whether a particular model and set of parameter choices is likely to have generated the data set.

Even with our focus on hypothesis testing, we still must consider parameter estimation. Many parameters describing HIV infection are known, at least within some range, but we are interested in escape rates precisely because these are unknown. Consequently, we describe a computational approach for forming a posterior distribution for the escape rates of different HIV variants. The point of these methods is to provide parameter values that can be used as a starting point for hypothesis testing.

While the model and inference methods we describe are general, in this work our results are restricted to the case of HIV escape at one or two epitopes. We first present numerical experiments showing that our posterior construction produces accurate estimates of escape rates and that our hypothesis tests have substantial statistical power. Then, we apply our methods to the data sets of patients CH40 and CH58 presented in Goonetilleke *et al.* (2009) and Fisher *et al.* (2010). Our methods use data that specify frequencies of different HIV variants at different sample timepoints, the type of data available in Goonetilleke *et al.* (2009) and Fisher *et al.* (2010).

Our analysis of CH40 focuses on the first, targeted epitope, meaning that we model escape at a single epitope. Using the deep-sequencing data set of Fisher *et al.* (2010), which provides variant frequencies soon after peak viral load, we construct a posterior for the escape rates of eight variants involved in the escape. The eight variants correspond to different mutations on the epitope, representing eight mutation pathways. Using the posterior as a guide, we test and reject the hypothesis that all eight variants share the same escape rate. In contrast, we are unable to reject a hypothesis that partitions the eight variants into three groups, within which escape rates are identical. The CH40 results demonstrate the utility of our methods in analyzing escape occurring through multiple mutation pathways.

Figure 3 Double-escape graph.

For CH58 we consider the first two regions at which mutations are seen to escape, meaning that we model escape at two epitopes. The two regions considered for CH58 are unrelated to the CH40 epitopes. The first region is a targeted epitope while the second region may be associated with fitness effects derived from changes in the targeted epitope region, but our techniques apply to any genome region under selection, regardless of whether the selection is CTL mediated. Using our posterior construction and hypothesis tests, we show that escape at the two regions may be intertwined, highly skewing escape-rate estimates that consider only a single epitope.

In our numerical experiments, as well as in considering the data sets CH40 and CH58, we compare our escape-rate estimates to those produced by fitting the standard model to data. Assuming that mutation and birth–death events are stochastic, we find that standard model-based estimates are downward biased while our estimates are roughly unbiased. More importantly, ignoring stochasticity severely underestimates the range of parameter values that can fit a data set, leading to very high type I errors.

Calculating posteriors and P -values under our model is not computationally trivial. The model is stochastic and the data are high dimensional. Standard Monte Carlo approaches in which the birth–death process is simulated without any conditioning are not effective because the data represent a single point in a high-dimensional space and the birth–death process rarely hits any specific point. Further, since the data are high dimensional, defining what is meant by a P -value is not straightforward. We address these issues by constructing an approximation of the birth–death process, which is a stochastically simpler process. Through this approximation, which we refer to as the stochastic reduction, posteriors can be computed using a Markov chain Monte Carlo approach and P -values can be defined and computed efficiently.

Importantly, our current hypothesis tests focus on deep-sequencing data sets. As described in the *Results* section, our algorithm for computing P -values is accurate when the num-

Table 1 Rates for the birth–death process

Cell type	Birth rate	Death rate	Mutation rate
T	λ	$d + \sum_v k l_v$	—
I_v	kT	$\delta_v(t)$	$\mu \sum_{v' \rightarrow v} k T l_{v'}$

ber of sampled sequences is large, ~ 5000 sequences. For CH40 we have access to deep-sequencing data, but for CH58 we do not. This is a general problem because deep-sequencing data are usually not linked across epitopes. However, as described more fully in the CH58 results section, we apply hypothesis tests to the CH58 data set by making assumptions on the variant frequencies part of the null hypothesis. Our posterior construction methods are valid regardless of sample size.

Model and Methods

We form our model by replacing the deterministic differential equations of the standard model, see (1) below, with stochastic equations that track the birth and death of individual infected cells. The transition from deterministic to stochastic equations is needed to model the rise of escape mutations, but this transition complicates inference methods. For the standard model, a given set of parameter values always produces the same dynamics. For our model, different simulations using identical parameters can produce significantly different dynamics.

Our methods are based on a reduction that replaces stochastic dynamics with samples from one-dimensional probability distributions. To explain this reduction, consider a single HIV variant v that initially infects a single CD4+ cell but that, over time, comes to infect many CD4+ cells. The v variant dynamics will initially be stochastic due to the randomness of birth and death events, but as time goes on and the number of v variants rises, averaging effects will reduce the stochasticity and deterministic equations will describe the dynamics. To achieve our reduction, we select a deterministic time, call it t , at which we are confident that the number of v infected cells will be sufficiently large to make the v dynamics deterministic. The number of v variants at t will be random, but we can calculate the underlying probability distribution. Our reduction involves replacing the stochastic dynamics of v variants through time, with a single sample from the probability distribution we compute at time t . Prior to t the number of v variants is small and can be roughly ignored in the context of the full infecting population; after t the v dynamics are roughly deterministic. In this way, we replace the stochastic model of v variant dynamics with a sample from a one-dimensional probability distribution and a deterministic model, a significant reduction in complexity. If we do this for all variants involved in an escape, we can reduce our stochastic model to a deterministic model along with a collection of samples from one-dimensional probability distributions.

Figure 4 Stochastic reduction for the escape graph of Figure 2. (A) Pop size distribution for M1 variants. (B) Pop size distribution for M12 variants. The solid (dashed) line is the M12 pop size distribution conditioned on M1 pop size of 1000 (3000). The M1 pop size is sampled from the M1 pop size distribution shown in subplot A. (C) Times for the start and end of the M12 stochastic interval, $[t_{\text{start}}^{(M12)}, t_{\text{end}}^{(M12)}]$.

Model

Our model is based on the following form of the standard model for which virions are assumed to be in steady state (Nowak and May 2000; Perelson 2002),

$$\begin{aligned}\dot{T}(t) &= \lambda - dT - kT \sum_v I_v \\ \dot{I}_v(t) &= kTI_v - \delta_v I_v,\end{aligned}\quad (1)$$

where T , I_v represent uninfected target CD4+ T cells and CD4+ T cells infected by HIV variant v , respectively, per microliter. To allow for multiple variants, v is an index over the set of variants considered. k , λ , d are the infection rate per target cell per microliter, target cell production rate per microliter, and target cell death rate, respectively, per day. Crucial for our model, δ_v is the death rate of cells infected by variant v . Instead of explicitly modeling CTL dynamics, we allow the parameters δ_v to vary in time, in this way modeling the selective force exerted by CTLs or other fitness effects implicitly.

To model the mutation pathways through which HIV escape occurs, we define an escape graph by specifying a set of vertices and directed edges. Vertices correspond to HIV variants that are part of the escape pathway. Two vertices, say A and B , may be connected by an edge directed from A to B if variant A can mutate into variant B through a single nucleotide substitution.

Although escape graphs can take any form, in this article we focus on single and double-escape graphs reflecting escape at one and two epitopes, respectively. Figure 1 is an example of a single-escape graph. Figure 2 and Figure 3 are examples of double-escape graphs. Below, we present results

involving Figure 1 and Figure 2, Figure 3 is given as an example demonstrating the generality of double-escape graphs. In all three figures, F represents the original founder HIV variant that infected the patient. In a single-escape graph, all mutants differ from the founder by a single nucleotide substitution on the targeted epitope and hence are directly connected to F by an arrow. In double-escape graphs, we have two epitopes and some variants have a nucleotide substitution at both epitopes, making their vertices two edges removed from F . For example, in the case of Figure 2, the movement from variant F to $M1$ to $M12$ represents a pathway of HIV escape in which single nucleotide substitutions transform F variants into $M1$ variants and then $M1$ variants into $M12$ variants. In this case, $M1$ represents a variant possessing a mutation on epitope 1 and $M12$ represents a variant possessing a mutation at both epitopes 1 and 2. The specific geometry of escape graphs, for example, Figure 2 lacks a variant that is mutated at epitope 2 by not 1, reflects a modeling choice.

A birth–death process describes the dynamics of the variant populations specified by the escape graph vertices and is an extension of (1). $T(t)$ and $I_v(t)$ represent the same populations as in the standard model, with the caveats that the v 's are vertices in the escape graph and the units are now per body rather than per microliter. The change in units allows us to track the rise of new variants, which initially infect only a single cell. The parameters of the birth–death process include the parameters of (1) with two extensions. First, we allow δ_v to depend on time; *i.e.*, we consider $\delta_v(t)$ instead of just δ_v . Second, if A and B are connected by an edge we let μ_{AB} be the rate at which A variants mutate into B variants. Throughout this work we set all such mutation rates equal

Figure 5 Hypothesis testing for the escape graph of Figure 2. (A and B) Pop size distribution for M1 variants and M12 variants. Solid black regions correspond to pop sizes >2200 (for A) and 4000 (for B). M12 pop size distribution assumes an M1 pop size of 2200 . (C and D) Same as plots A and B except that the M1 and M12 pop sizes are 1200 and 800 , respectively.

to $\mu = 3 \times 10^{-5}$ (Mansky 1996), but the methods allow for any value for any edge. The birth–death process is defined through the birth and death rates specified in Table 1. For example, at time t a single cell infected by variant v produces child infected cells with rate $kT(t)$, meaning that in a small time interval $[t, t + \Delta t]$ the probability of a new v variant infected cell arising is roughly $kT(t)I_v(t)\Delta t$. In all the cases we consider in this article, the escape graph includes a founder vertex F , but the methods allow for infection by multiple founder variants. We always start the birth–death process at “initial infection,” which we model as $t = 0$, $I_F(0) = 1$ and $I_v(0) = 0$ for $v \neq F$.

Since we do not explicitly model CTL dynamics, the $\delta_v(t)$ that implicitly model CTL attack are parameters of special interest. In computing posteriors, estimating $\delta_v(t)$ with no restrictions raises identifiability issues; see below. Instead, along with the escape graph and birth–death rates, we specify a list of attack intervals $[0, t_1]$, $[t_1, t_2]$, $\dots$, $[t_{n-1}, t_n]$ and restrict the $\delta_v(t)$ to be constant during any attack interval. For example, we often choose $t_1 = 15$ reflecting ~ 15 days before CTL response arises and correspondingly choose $\delta_v(t) = 0.4$ for $t \in [0, 15]$, which gives a half-life for infected cells falling between 1 and 2 days (Perelson *et al.* 1996; Stafford *et al.* 2000).

As mentioned, in this article we focus on data sets reflective of single and double epitope escape. Correspondingly, we consider two choices for the attack intervals. In what we refer to as the single-escape model, we assume a single-escape graph and two attack intervals: $[0, t_A]$, $[t_A, t_D]$. t_A , t_D are the day post-infection at which CTL response begins and the data were sampled, respectively. For each variant v in the escape graph, the model includes the death rate parameters $\delta_{v,1}$ and $\delta_{v,2}$ for times in $[0, t_A]$ and $[t_A, t_D]$ respectively. In what we refer to as the double-escape model, we assume a double-escape graph and three attack intervals: $[0, t_A]$, $[t_A, t_{D1}]$,

$[t_{D1}, t_{D2}]$, where t_{D1} , t_{D2} are times of two data samples. t_{D1} is a time appropriate for a single-escape model, *i.e.*, a time that captures the first CTL response and epitope escape, while t_{D2} is a time at which a second epitope has drawn significant CTL response and the sequence data show epitope mutations at both epitopes targeted. In the double-escape model, each variant is associated with three death rates: $\delta_{v,1}$, $\delta_{v,2}$, and $\delta_{v,3}$ corresponding to the three attack intervals.

For our single-escape model, an appropriate data set samples sequences at a time when the first targeted epitope and mutations at that epitope are both at significant frequencies, but before mutations at other targeted epitopes arise. Such a sampling time allows us to capture a variety of epitope mutations, while still focusing on escape at a single epitope. CH40 is an example of such a data set. An appropriate data set for our double-escape model is, essentially, an appropriate data set for a single-escape model with the addition of a later sampling timepoint that capture escape at the second targeted epitope. At the second sampling time, epitope mutations at the second targeted epitope should be at significant frequency, while mutations at epitopes beyond the first two epitopes targeted should be of low frequency to focus on double escape. CH58 is an example of such a data set.

Our main interest lies in the parameters $\delta_{v,2}$ for the single-escape model and $\delta_{v,2}$, $\delta_{v,3}$ for the double-escape model, where v varies over all vertices in the escape graph. However, two identifiability issues arise in the context of our models. We refer the reader to Miao *et al.* (2011) for a review of identifiability, but for us identifiability roughly means that we can uniquely infer death rates from the data if we ignore stochastic effects.

The first identifiability problem we face arises from the form of our data and applies to both single- and double-escape models. Roughly, in the context of the single-escape model, shifting all the $\delta_{v,2}$ by an identical amount, say to

Figure 6 Relative error for escape-rate estimates under the single-escape model. The numbers 1, 3, 5, 8 give the number of epitope mutations involved in the escape. For each such number, the left (right) boxplot gives the absolute error scaled by 0.4 produced by the posterior (deterministic) estimation method.

$\delta_{v,2} + c$ for some constant c , does not change the frequencies of the variants at a particular timepoint. This means frequency data cannot be used to estimate all the $\delta_{v,2}$, even in the absence of stochastic effects. (See [Supporting Information, File S1](#), for more details.) With this in mind, we estimate the differences $\delta_{F,2} - \delta_{v,2}$ and $\delta_{F,3} - \delta_{v,3}$ where now v are all vertices other than the founder vertex. This same identifiability issue exists for existing techniques; see discussion in Asquith *et al.* (2006).

Following previous authors (Fernandez *et al.* 2005; Asquith *et al.* 2006; Goonetilleke *et al.* 2009; Ganusov *et al.* 2011), we refer to $\delta_{F,2} - \delta_{v,2}$ and $\delta_{F,3} - \delta_{v,3}$ as the first- and second-escape rates of variant v . For a single-escape model, a variant has only a single-escape rate, $\delta_{F,2} - \delta_{v,2}$.

The second identifiability problem arises in the context of double-escape model data. As an example, consider the escape graph of Figure 2. In many data sets that describe double or a higher number of epitope escapes, the first sampled timepoint taken soon after peak viral load reveals only mutations on the first epitope targeted. Consequently, the first timepoint provides no information about the death rates of variants with mutations at other epitopes. In our notation, for the double-escape model, this means that $\delta_{v,2}$ and $\delta_{v,3}$ for v representing a variant with mutations outside the first targeted epitope cannot simultaneously be identified. In such cases we choose a value for $\delta_{v,2}$ and use our hypothesis tests to verify that such values are reasonable.

Stochastic reduction

Since the number of infected cells in acute infection is of order 10^9 , simulating our birth–death process through a Gillespie algorithm (Gillespie 2001) is not computationally feasible. Previous authors, who have considered birth–death processes similar to ours, have used the idea of a stochastic-deterministic decomposition to construct simulations (Nowak 2000; Rouzine *et al.* 2001; Desai and Fisher 2007; Leviyang 2012). In a stochastic-deterministic decomposition, the number of cells infected by a given variant v are simulated using a Gillespie algorithm until they reach some large cutoff value; after that time a deterministic differential equation is used to describe v variant dynamics.

Our stochastic reduction is also based on a stochastic-deterministic decomposition, but with the aim of expressing

the birth–death process in a form more amenable to inference. To do this, for each variant v we define a stochastic interval, $[t_{\text{start}}^{(v)}, t_{\text{end}}^{(v)}]$. Given the stochastic interval, $I_v(t)$ dynamics are generated according to the following algorithm.

Prestochastic interval step:

$$\text{For } t < t_{\text{start}}^{(v)}, I_v(t) = 0.$$

Stochastic interval step, substep a: Set $I_v(t) = 0$ for all $t \in [t_{\text{start}}^{(v)}, t_{\text{end}}^{(v)}]$. During this time interval, store $T(t)$ and $I_{v'}(t)$ for all v' with an edge pointing to v .

Stochastic interval step, substep b: Let $B(t)$ be a single-variant birth–death process; *i.e.*, $B(t)$ is a scalar, defined by $B(t_{\text{start}}^{(v)}) = 0$ and with birth, death, and mutation rates of $kT(t)$, $\delta_v(t)$, and $\sum_{v' \rightarrow v} \mu kT(t)I_{v'}(t)$, respectively. $B(t)$ has the same birth, death, and mutation rates as variant v infected cells. Define $X_v = B(t_{\text{end}}^{(v)})$. Since $T(t)$, $I_{v'}(t)$ were stored in substep a, the distribution of X_v can be computed through standard methods. See [File S1](#), for full details, but briefly, X_v is computed by solving a backward equation for the expected value $E[\exp[i\omega B(t_{\text{end}}^{(v)})] | B(t) = 1]$. Then a Fourier transform in ω is performed to obtain the distribution of $B(t_{\text{end}}^{(v)})$ or, in other words, X_v . See also Conway and Coombs (2011) for a similar computation in the context of viral dynamics.

Stochastic interval step, substep c: Using the distribution computed in substep b, produce a sample $\hat{x}_v$ from X_v . Set $I_v(t_{\text{end}}^{(v)}) = \hat{x}_v$.

Deterministic interval step: For $t > t_{\text{end}}^{(v)}$, $I_v(t)$ dynamics are approximated using the differential equation:

$$\dot{I}_v = kTI_v - \delta_v(t)I_v + \mu \sum_{v' \rightarrow v} kTI_{v'}. \quad (2)$$

The above algorithm can be employed for each variant, but in practice we assume that the founder variant is deterministic from $t = 0$. This approach ignores stochasticity in the first 1–2 days of infection, a time period that does not affect our results.

Figure 7 Comparison of escape-rate methods applied to the same data set. As shown, the actual escape rate was 0.27 as shown in the far left. The posterior distribution is relatively wild and contains the correct escape-rate value. In contrast the deterministic distribution is relatively narrow and does not contain the correct escape-rate value.

The endpoints of the stochastic intervals, $[t_{\text{start}}^{(v)}, t_{\text{end}}^{(v)}]$, are set through tuning parameters ε and L . These two parameters tune the algorithm by implicitly selecting a trade-off between computational speed and accuracy. To explain how we choose $t_{\text{start}}^{(v)}$ and $t_{\text{end}}^{(v)}$, we consider the escape graph in Figure 2. Starting with variant M1, since F does not have a stochastic interval, $t_{\text{start}}^{(M1)}$ is defined by $\mu k T(t_{\text{start}}^{(M1)}) I_F(t_{\text{start}}^{(M1)}) = \varepsilon$, i.e., the time at which the rate of mutations reaches ε . $t_{\text{end}}^{(M1)}$ is defined by $\tilde{I}_{M1}(t_{\text{end}}^{(M1)}) = L$, where $\tilde{I}_{M1}(t)$ is roughly the average population size of M1 variants. Put precisely, let t^* be defined by $\mu k I_F(t^*) T(t^*) = 1$; then $\tilde{I}_{M1}(t)$ obeys (2) with initial condition $\tilde{I}(t^*) = 0$. The stochastic interval of M12 is generated similarly, but with M1 playing the role of F . For a general escape graph, we would start with the founder and work our way outward along the edges, generating a stochastic interval for each vertex.

The above algorithm introduces errors in the simulation of the birth–death process in each of the three steps. In the prestochastic step, we ignore the order ε probability that mutations create v variants prior to $t_{\text{start}}^{(v)}$. In the stochastic interval step we assume that $I_v(t) = 0$ even though v variants increase from 0 to order L . In the deterministic interval step, we ignore stochastic effects of the birth–death process and assume deterministic dynamics.

The stochastic-deterministic decomposition used by previous authors to simulate birth–death processes is obtained by replacing our stochastic interval, substeps a–c by a Gillespie simulation. These previous works show that the relative error produced by the stochastic interval and deterministic interval steps is < 0.03 and 0.01 for $L = 100$ and $L = 2000$, respectively (Rouzine *et al.* 2001; Desai and Fisher 2007; Leviyang 2012). Importantly, as emphasized in Rouzine *et al.* (2001), the accuracy of the deterministic interval step depends on relatively strong selection, a condition satisfied in acute HIV infection. In our results below, we take $L = 2000$. The error associated with the prestochastic step is order ε . (See File S1 for further discussion of these errors.)

Using the algorithm above, the stochasticity of the birth–death process is reduced to the variables $I_v(t_{\text{end}}^{(v)})$ or, in other words, the draws of the samples $\hat{x}_v$ from the distributions X_v . To see this, note that except for sampling from $\hat{x}_v$ for all variants $v \neq F$, v variant dynamics are completely determin-

istic. We refer to X_v and $\hat{x}_v$ as the pop size distribution and pop size of the v variant population because, intuitively, $\hat{x}_v$ determines how soon the v variant population has significant frequency and hence “pops up” in the data.

At this point we have already achieved a reduction by expressing the stochasticity of the birth–death process in terms of the pop size distributions. However, for appropriate choices of ε and L the stochasticity of the birth–death process can be further simplified. For concreteness, consider the escape graph of Figure 2. If ε is chosen large enough and L is chosen small enough then $t_{\text{end}}^{(M1)} < t_{\text{start}}^{(M12)}$; i.e., the stochastic interval of M1 will end before the stochastic interval of M12 starts. In this case, the stochastic interval of M1 will be considered first, the distribution of X_{M1} generated, and $\hat{x}_{M1}$ sampled. Only once $\hat{x}_{M1}$ has been sampled will the dynamics be run forward to the M12 stochastic interval. Then, X_{M12} will be generated and $\hat{x}_{M12}$ will be sampled. For more complex escape graphs, the idea is the same. The pop size of a vertex v depends on the pop sizes sampled for vertices that have an edge pointing to it and sampling of pop sizes on the escape graph begins from the founder vertex and moves outward.

To make these comments concrete, consider again the escape graph of Figure 2. Figure 4 shows the pop size distributions for M1 and M12 variants (see File S1 for parameter values used). The M1 stochastic interval ends at $\sim t = 11$. Plot A gives the pop size distribution of M1, i.e., the distribution of the number of M1 variants at $t = 11$. Under our algorithm, instead of simulating the dynamics of M1 variants, we set the number of M1 variants at $t = 11$ by sampling from this distribution. Once a pop size has been chosen for M1, the M12 stochastic interval and pop size distribution can be determined. Plot B shows the pop size distribution of M12 for two values of the M1 pop size, 1000 and 3000. Plot C shows the start and end time of the M12 stochastic interval for different values of the M1 pop size. As in the case of M1 variants, at the end of the M12 stochastic interval, we sample from the M12 pop size distribution to determine the number of M12 variants.

Computing posteriors

Let $S(t)$ be the state at time t of the populations tracked by the birth–death process, e.g., for the escape graph of Figure 2, $S(t) = (T(t), I_F(t), I_{M1}(t), I_{M12}(t))$. Let $\hat{D}(t_1, t_2)$

Table 2 Type I error estimates: single-escape model

Sample size	Confidence level (%)	No. of mutations			
		1	3	5	8
∞	99	0.01	0.01	0.01	0.01
	95	0.05	0.06	0.04	0.05
	90	0.10	0.11	0.10	0.11
	80	0.21	0.21	0.20	0.24
	50	0.52	0.52	0.53	0.54
5000	99	0.03	0.04	0.06	0.07
	95	0.09	0.11	0.14	0.15
	90	0.14	0.18	0.21	0.25
	80	0.25	0.29	0.33	0.37
	50	0.54	0.56	0.60	0.64
1000	99	0.09	0.18	0.26	0.34
	95	0.14	0.29	0.39	0.49
	90	0.20	0.35	0.47	0.58
	80	0.29	0.44	0.57	0.68
	50	0.55	0.65	0.76	0.84

be frequency data for the variants of the escape graph collected at timepoints t_1, t_2 . We choose two timepoints, which correspond to a double-escape model; for concreteness, in general, any number of data timepoints are possible. As an example, the CH58 data discussed in *Results* include two timepoints; see Table 6. Let θ represent the parameters of the birth–death process for which we want to build a posterior; throughout this article θ is composed of escape rates, but in general any set of parameters can be chosen. We can simulate the birth–death process and generate samples for $S(t_1), S(t_2)$. Let $\mathcal{D}(S(t_1), S(t_2))$ be the data generated by simulating the birth–death process and then simulating the sampling of sequences. By this we mean that first, through simulation, $S(t_1), S(t_2)$ must be sampled to establish the exact frequencies of the variants at times t_1, t_2 . Then, hypothetical samples must be drawn at time t_1 and t_2 to form simulated data.

Given a prior for θ , $\pi(\theta)$, our goal is to compute a posterior of θ conditioned on the data. More precisely, we aim to compute

$$P(\theta | \mathcal{D}(S(t_1), S(t_2)) = \hat{\mathcal{D}}(t_1, t_2)). \quad (3)$$

However, it is easier to compute a posterior for $\theta, S(t_1), S(t_2)$,

$$P(\theta, S(t_1), S(t_2) | \mathcal{D}(S(t_1), S(t_2)) = \hat{\mathcal{D}}(t_1, t_2)), \quad (4)$$

and (3) can be obtained from (4) by treating $S(t_1)$ and $S(t_2)$ as nuisance parameters.

Bayesian posteriors are often computed through Markov chain Monte Carlo (MCMC) methods; see Lemey *et al.* (2009, Chap. 7) for a review of MCMC theory applied to viral data and Gilks *et al.* (1996) for a general review. In our context, implementing such an approach depends on being able to compute (4). Once (4) can be computed, various MCMC methods allow one to sample from the posterior of θ . Specifically, we implement a Metropolis–Hastings-based

MCMC. Such an approach is not affected if instead of computing (4) we compute

$$P(\theta, S(t_1), S(t_2), \mathcal{D}(S(t_1), S(t_2)) = \hat{\mathcal{D}}(t_1, t_2)). \quad (5)$$

Indeed, (5) is identical to (4) up to a constant factor and such a proportional factor has no effect on a Metropolis–Hastings MCMC.

Equation 5 can be expressed as a product of simpler conditional probabilities,

$$\begin{aligned} P(\theta, S(t_1), S(t_2), \mathcal{D}(S(t_1), S(t_2)) = \hat{\mathcal{D}}(t_1, t_2)) \\ = P(\mathcal{D}(S(t_1), S(t_2)) = \hat{\mathcal{D}}(t_1, t_2) | S(t_1), S(t_2), \theta) \\ \times P(S(t_1), S(t_2) | \theta) \pi(\theta). \end{aligned} \quad (6)$$

The factor $P(\mathcal{D}(S(t_1), S(t_2)) = \hat{\mathcal{D}}(t_1, t_2) | S(t_1), S(t_2), \theta)$ can be interpreted as a sampling probability, that is, conditioned on knowing $S(t_1), S(t_2)$ and hence the frequencies of the variants at times t_1, t_2 , what is the probability of drawing the data. Computation of such probabilities is standard. However, $P(S(t_1), S(t_2) | \theta)$, the probability of a given system state at t_1, t_2 given a parameter choice, is not standard.

Our approach is to use the stochastic reduction to replace $S(t_1), S(t_2)$ by $\hat{x}_v$ for all variants $v \neq F$. Since the birth–death process under our approximation is completely determined by the pop size samples, the $\hat{x}_v$ determine the $S(t)$. Given this equivalence, we consider $P(\hat{x}_v \text{ for } v \neq F | \theta)$ instead of $P(S(t_1), S(t_2) | \theta)$. Since we know the distribution of each $\hat{x}_v$ and since the dependence structure of the $\hat{x}_v$ is simple by the stochastic reduction, $P(\hat{x}_v \text{ for } v \neq F | \theta)$ can be computed in a standard manner.

We use a random walk Metropolis–Hastings algorithm on θ and the $\hat{x}_v$ to form the posterior. For instance, to form the posterior for patient CH58, θ is three dimensional and the $\hat{x}_v$, namely $\hat{x}_{M1}$ and $\hat{x}_{M12}$, are two dimensional. Our MCMC then operates on a five-dimensional state space. To compute 10^7 MCMC steps takes ~ 1 day on an Intel I7-2600 using our C++ implementation. In all our results, we use a uniform prior on $[-1, 2]$ for each escape rate.

Hypothesis testing

Given an escape graph and a choice of parameters for the birth–death process, our goal is to test the null hypothesis that the data are formed by the model. Here we let θ represent all the parameters of the birth–death process and the underlying escape graph, as opposed to the previous section where θ typically represents only the escape rates. The challenge lies in computing a P -value. More precisely, our goal is to compute the P -value of the data, $\hat{\mathcal{D}}(t_1, t_2)$, given θ . Given a P -value, we can then test the null hypothesis that the data came from a model with parameters θ . Since the data are multidimensional, forming a statistic through which to compute a P -value is not straightforward. However, as in the case of posterior computations, the stochastic reduction allows for a simplification.

Table 3 Power estimates: single-escape model

No. mutations	Distance to null model						
	0.01–0.03	0.03–0.06	0.06–0.09	0.09–0.12	0.12–0.15	0.15–0.18	0.18–0.21
1	0.35	0.80	0.96	0.99	1	1	1
3	0.38	0.81	0.94	0.98	0.98	0.99	1
5	0.40	0.84	0.97	0.99	1	1	1
8	0.36	0.80	0.95	0.99	1	1	1

Using the escape graph of Figure 2 as a concrete case, for a given θ there will be a pair of pop sizes $\hat{x}_{M1}, \hat{x}_{M12}$ for which $P(\mathcal{D}(\hat{x}_{M1}, \hat{x}_{M12}) = \hat{D}(t_1, t_2) | \theta)$ is maximized, here we are replacing $\mathcal{D}(S(t_1), S(t_2))$ by $\mathcal{D}(\hat{x}_{M1}, \hat{x}_{M12})$ since the pop sizes completely determine the dynamics of the birth–death process. Label the $\hat{x}_{M1}, \hat{x}_{M12}$ that achieve the maximum as $x_{M1}^{\text{data}}, x_{M12}^{\text{data}}$. Then we can assess the P -value of $\hat{D}(t_1, t_2)$ by considering where $(x_{M1}^{\text{data}}, x_{M12}^{\text{data}})$ fall in the densities of X_{M1}, X_{M12} .

Before making this approach mathematically precise, we give a concrete example. Consider the example discussed at the end of the *Stochastic Reduction* subsection involving the escape graph in Figure 2. All model parameters have the same values mentioned in that case. Now, however, we bring in data. Suppose that we have a one-timepoint data set for which at $t = 40$ the frequencies for variants $F, M1, M12$ are 0.05, 0.6, 0.35, respectively (since Figure 2 is a double-escape graph, we would usually have two timepoints, but here we use one for simplicity of presentation). Standard numerical optimization on the two-dimensional space composed of the $M1$ and $M12$ pop sizes shows that these frequencies are achieved when the $M1, M12$ pop sizes are $\sim 2200, 4000$, respectively. In the notation discussed above, $x_{M1}^{\text{data}} = 2200$ and $x_{M12}^{\text{data}} = 4000$. The $M1$ pop size distribution depends only on the model parameters, which are fixed, while the $M12$ pop size distribution depends on the $M1$ pop size, which we take as 2200. We can then ask where the values 2200 and 4000 fall in these distributions. Figure 5, A and B, gives the $M1$ and $M12$ pop distributions; the black region corresponds to pop sizes that are >2200 and 4000, respectively. More precisely, the black regions give us the P -values 0.11 and 0.05 for the 2200 and 4000, respectively. Combining these P -values, we compute an overall P -value of 0.04, meaning that at a confidence level of 0.95 the null hypothesis that the model formed the data can be rejected.

Now we consider a slightly different data set. Suppose that the data set is as before, except that now we have frequencies of 0.1, 0.8, 0.1 for $F, M1, M12$, respectively. Importantly, we are working with the same model parameters as in the preceding paragraph. Under these data, we now find the optimal $M1$ and $M12$ pop sizes to be 1200 and 800, respectively. Figure 5, C and D, shows where these pop sizes fall on the pop size distributions. We find P -values of 0.42 and 0.19 for 1200 and 800, respectively. Combining these P -values, we find an overall P -value of 0.29, meaning that we cannot reject the null hypothesis that the model formed the data.

To make the approach precise, let Φ_{M1} and Φ_{M12} be the cumulative densities of X_{M1} and X_{M12} respectively. Then let $\Phi_{M1}(x_{M1}^{\text{data}}) = q_{M1}$ and $\Phi_{M12}(x_{M12}^{\text{data}}) = q_{M12}$; i.e., q_{M1}, q_{M12} are the quantiles of $x_{M1}^{\text{data}}, x_{M12}^{\text{data}}$. If x_{M1}^{data} was chosen from the density of X_{M1} , q_{M1} would be uniformly distributed on $[0, 1]$. The same holds for q_{M12} . Further, due to the stochastic reduction, q_{M1} and q_{M12} are independent because the sampling from the the pop size distributions X_{M1} and X_{M12} occurs independently even though the distribution X_{M12} depends on X_{M1} . We can then define the P -value by computing the probability with which two independent, uniform random numbers are more “extreme” than q_{M1}, q_{M12} . This is a standard statistical problem; see [File S1](#) for the specific algorithm we use to compute the P -value from the q ’s.

As noted in the preceding paragraph, q_{M1} and q_{M12} are uniformly distributed on $[0, 1]$ if $x_{M1}^{\text{data}}, x_{M12}^{\text{data}}$ are assumed drawn from the pop size distributions, i.e., X_{M1} and X_{M12} , respectively. This, however, is not precisely true. To see this, imagine a scenario in which we simulate our model and construct a data set; indeed, the null hypothesis assumes that the data set is generated in exactly this way. The simulation determines variant frequencies at time t_1, t_2 corresponding to the pop sizes sampled, $\hat{x}_{M1}, \hat{x}_{M12}$. Using these variant frequencies we construct a data set $\hat{D}(t_1, t_2)$ by sampling sequences. Importantly, due to sampling error, the variant frequencies in the data set may not be the same as those produced by the simulation. Since we compute $x_{M1}^{\text{data}}, x_{M12}^{\text{data}}$ using the data set, $x_{M1}^{\text{data}}, x_{M12}^{\text{data}}$ will not equal $\hat{x}_{M1}, \hat{x}_{M12}$, meaning that $x_{M1}^{\text{data}}, x_{M12}^{\text{data}}$ are not samples from the pop size distributions.

As discussed in *Results*, using the pop size distributions for the distributions of $x_{M1}^{\text{data}}, x_{M12}^{\text{data}}$ is an approximation that is accurate for large sample sizes. In the case of small sample sizes, we could compute the distribution of $x_{M1}^{\text{data}}, x_{M12}^{\text{data}}$ rather than using the pop size distributions as approximations, but such an approach is computationally expensive and requires further work. We can apply our current P -value computations to data sets with small sample size by including frequency assumptions in the null hypothesis; see CH58 results for a concrete example.

Results

We first present numerical experiments exploring our inference methods under a range of parameter values. Then we turn to patient data, applying our methods to the CH40 and CH58 data sets. We examine the CH40 data set through

Table 4 CH40 data at day 16

Label	a.a. sequence	Frequency at day 16 (%)
<i>F</i>	SLAFRHVAR	50
<i>Q</i>	—Q—	4.7
<i>H</i>	—H—	24.9
<i>N</i>	N—	1.9
<i>R</i>	R—	3.8
<i>M</i>	—M—	5.6
<i>E</i>	—E—	2.0
<i>C</i>	—C—	5.7
<i>I</i>	I—	1.4

the single-escape graph, while considering CH58 through the double-escape graph.

Numerical experiments

Through numerical experiments, we investigated the error of our parameter inference methods as well as the type I error and statistical power of our hypothesis testing methods. Importantly, the experiments show that inference based on the standard model produces overly narrow confidence intervals as a consequence of ignoring stochastic effects, most notably mutation. In contrast, our methods produce more accurate confidence intervals.

We present numerical results for single-escape graphs. Results for double-escape graphs are similar; see [File S1](#). Single-escape graphs have the form of Figure 1 with only the number of vertices pointed to by *F* varying. We consider single-escape graphs with 1, 3, 5, and 8 epitope mutations. Recall that in the single-escape model, $\delta_{v,2}$ represents the death rate of variant *v* under CTL response. We focus on inferring the escape rates $\delta_{F,2} - \delta_{v,2}$ over all mutant variants *v*.

Parameter inference experiments: Our first numerical result shows that our posterior construction methods provide accurate estimates for the escape rates. We conducted a numerical experiment consisting of the following steps:

1. We randomly chose parameter values for our model from a meaningful biological range.
2. Given a specific choice for the parameter values, we simulated the resulting model and produced a simulated data set of variant frequencies at time t_D .
3. Given the simulated data set, we constructed a posterior for the escape rates, assuming no knowledge of the death rates, but assuming we knew other parameters of the model.

To compare our posterior method to an approach dependent on a deterministic model, we performed an experiment using steps 1 and 2 as above. But for step 3, we used a least-squares approach to fit a standard model to the simulated data set. Importantly, sampling stochasticity at time t_D was preserved, allowing us to form a distribution of estimates and corresponding confidence intervals under the standard

model. See [File S1](#) for more specifics on both posterior and deterministic experiments.

We repeated the above experiments 1000 times for both the posterior and deterministic inference methods assuming 5000 samples at time t_D . Figure 6 gives a boxplot for the escape-rate error using posterior and deterministic methods distributions. We present the error scaled by 0.4 since that is approximately the death rate of HIV infected cells in the absence of CTL attack (Perelson *et al.* 1996), so 0.4 can be seen as the background noise through which CTL influence on death rates must be estimated. For a given number of epitope mutations, the left and right boxplot corresponds to the posterior and deterministic method, respectively. The results show that the constructed posteriors are roughly unbiased with relative errors usually <0.05 . The deterministic estimators are biased down because mutations start immediately in the deterministic model, while in the stochastic model founder variants must first rise to significant numbers. Since the number of mutations is overestimated by the deterministic method, the escape rates must be lower to achieve the same variant frequencies.

Figure 6 does not give an entirely accurate picture because the errors are averaged over 1000 simulated data sets. Figure 7 was generated from a three-epitope mutation model; see [File S1](#) for details. The figure compares the estimation of one of the three escape rates through posterior and deterministic methods. In contrast to Figure 6, the results represent estimates rather than scaled errors. As shown, the actual escape rate was 0.27. Posterior construction produces a relatively wide distribution containing the correct escape rate. Since the deterministic estimates ignore the stochasticity of mutation, considering only sampling stochasticity, the deterministic distribution is too narrow. This case, which is typical, is reflected in the poor type I errors discussed below.

Hypothesis testing: We consider two null hypotheses. For our methods, the null hypothesis assumes that the simulated data set was generated under our stochastic model with the parameter values chosen. In contrast, to examine deterministic, standard model based methods, the null hypothesis assumes that the simulated data set was generated under the standard model with the parameter values chosen. To examine type I error under both methods, we chose parameters and simulated data sets exactly as in steps 1 and 2 of the experiments described above. Then, the null hypothesis was evaluated by computing *P*-values. When assuming our stochastic model, the *P*-value accounts for stochastic effects derived from both the model and sampling at time t_D . In contrast, when assuming a standard model, the *P*-value reflects solely sampling stochasticity (see comments in [File S1](#) involving confidence intervals under the standard model). Type I error estimates given below were produced by averaging over 10,000 hypothesis tests.

Type I errors for the deterministic method were very high, reflecting the overly narrow distributions constructed

Figure 8 Posterior and deterministic estimates for escape rates of CH40 epitope mutations. Each tic labeled by a letter and its right neighboring tic represent the posterior and deterministic estimates, respectively, for the corresponding mutation.

by the deterministic method. For example, in the case of 5 epitope mutations and 5000 samples at t_D , type I error was 0.567, 0.65, 0.76 for confidence levels of 99, 95, and 80%, respectively.

Table 2 shows the type I error for our methods at different confidence levels and number of sequences sampled. The sample size of ∞ reflects exact knowledge of the variant frequencies at sampling time, *i.e.*, no sampling error. Allowing for error associated with running 10,000 experiments, Table 2 shows that a sample size of ∞ has the correct type I error for the various confidence levels. As seen in the table, type I errors rise as the sample size drops, so that our hypothesis methods are too conservative in accepting the null hypothesis for finite sample sizes. The table shows that type I errors increase as the number of epitope mutations rises. Note that even at the sample size of 1000, our type I errors are more accurate than those obtained by deterministic methods under a sample size of 5000.

Two factors influence the type I errors at finite sample sizes. First, when a mutation variant is of low frequency, sampling errors become more pronounced and type I errors increase. For example, consider the case of 1 mutant and sample size of 1000. With unrestricted variant frequencies the table shows type I errors of 0.09 and 0.14 for confidence levels of 99 and 95%, respectively. However, if we consider only those experiments for which the mutation variant has frequency >0.05 at sampling time, we find type I errors of 0.06 and 0.10, respectively. Second, to understand why coverage error rises as the number of mutants rises, recall that we compute a P -value for each mutant and then combine all such P -values to produce an overall P -value. We find that regardless of the number of mutants, our P -value estimates for each individual mutant have roughly the same accuracy. However, when there are more mutants, the error in estimating the P -value of each mutant accumulate and produce a less accurate overall P -value. Further, as the number of mutants rises, there are often more mutants at low frequency, leading to the error described above.

To examine power, we again chose parameters and simulated data sets exactly as we did for parameter inference. However, we then constructed another set of parameter choices as follows. For each $\delta_{v,2}$ we defined $\tilde{\delta}_{v,2} = \delta_{v,2} + u_v$, where the u_v ranged on $[0, 0.3]$. We did the same for $\delta_{F,2}$. For

this new collection of parameters, we performed a hypothesis test using the simulated data set. Note that the data set is simulated using the $\delta_{v,2}$, while the $\tilde{\delta}_{v,2}$ are our null model. We repeated this experiment 10,000 times.

Given the poor type I errors of the deterministic method, we present power results only for our hypothesis-testing methods. Table 3 shows the power as a function of the Euclidean distance between the $\delta_{v,2}$ and $\tilde{\delta}_{v,2}$ at the confidence level 95% assuming 5000 sampled sequences. As can be seen, we are able to reject appropriately when the true death rates are ~ 0.1 apart from the null hypothesis death rates. Power under 1000 sampled sequences is slightly greater (data not shown), but this increase comes with the price of greater type I error as noted in Table 2.

Patient CH40

The data we use for patient CH40 is presented in Goonetilleke *et al.* (2009) and Fisher *et al.* (2010). We refer the reader to those articles and references therein for full details. Briefly, patient CH40 was identified during acute infection. Viral load data and immune response assays suggest the time of identification to be a few days after peak viral load. In Goonetilleke *et al.* (2009), identification time was labeled as day 0 (note that day 0 is not the time of initial infection). Blood samples were collected at day 0, day 16, and subsequent times. Through single-genome amplification and T-cell assays, Goonetilleke *et al.* (2009) determined that initial HIV escape occurred at the epitope NEF SR9 and that NEF SR9 was indeed recognized by T cells. Escape at NEF SR9 was first seen in the day 16 samples, meaning 16 days after identification. Day 0 samples were homogeneous for the epitope. Other epitopes and CTL responses were seen, but only at later sampling times (see Figure 2, component CH40.t in Goonetilleke *et al.* 2009). Fisher *et al.* (2010) used deep-sequencing methods on CH40 day 16 samples recovering 7754 sequences containing the NEF SR9 region. Table 4 gives the NEF SR9 epitope, the epitope mutations, and the variant frequencies at the day 16 sampling. The table is almost identical to data found in Fisher *et al.* (2010, Figure 5) except that we consider only variants with frequency $>1\%$ at day 16. For those variants, frequencies were rounded to the nearest tenth and rescaled, by a factor of 1.033, so that frequencies summed to 100%. Deep-sequencing data may have high error rates for

Table 5 CH40 parameters

Variant	δ during [0, 15]	δ during [15, 36]
F	0.4	0.8
Mutants	0.4	$0.8 - \delta_M$

low frequency variants. The cutoff we use of 1% may be too low. Here we do not focus on this important issue, assuming that sampling error for deep-sequencing data follows standard multinomial distributions. As specified in the table, each NEF SR9 mutant is associated with a one-letter label, the amino acid substitution caused by a single nucleotide mutation.

We used a single-escape model to investigate HIV escape at NEF SR9. Recall that single-escape models involve two time intervals, $[0, t_A]$ and $[t_A, t_D]$ representing time periods before and during CTL response, respectively, measured in days post-infection (not identification). We took $t_A = 15$ and $t_D = 36$, meaning that we assume day 16 post-identification is day 36 post-infection. The escape graph is given in Figure 1. The standard model parameters were chosen to fit the viral load pattern of patient CH40, see Goonetilleke *et al.* (2009, Figure S1), and to fall within known ranges (Perelson *et al.* 1993; Stafford *et al.* 2000; De Boer 2007). Specifically, $k = 3 \times 10^{-3}$, $d = 0.01$, $\lambda = 10^8$. We also set $\delta_{v,1} = 0.4$ for all variants in the escape graph; recall that $\delta_{v,1}$ is the death rate of variant v during $[0, t_A]$. This choice reflects an assumption of equal fitness prior to CTL response. Our goal was to infer the escape rates, $\delta_{F,2} - \delta_{v,2}$.

We first constructed posterior and deterministic distributions to estimate the escape rates (deterministic estimates were constructed as described in the numerical experiments). Figure 8 gives the marginal distributions for each of the epitope mutations. As in Figure 7, the deterministic distribution is biased down.

We used the posterior estimates as a basis for a hypothesis test. Specifically, we used the median values found in Figure 8 as values for $\delta_{F,2} - \delta_{v,2}$. Following from the identifiability issues mentioned, the choice we make for $\delta_{F,2}$ has no effect on the P -value so we arbitrarily chose $\delta_{F,2} = 0.8$. Using the resulting parameter values for $\delta_{v,2}$ gave a P -value of 0.99, demonstrating that the posterior does provide parameters that fit the data well.

Ganusov *et al.* (2011) used existing methods, assuming a single mutation variant to infer an NEF SR9 escape rate of 0.22 with a 95% confidence interval of 0.11–0.536. Importantly, they used data from Goonetilleke *et al.* (2009), which contain 13 sequences rather than the 7754 of the Fisher *et al.* (2010) data set. Further, since current approaches require two sample timepoints, they estimated escape rates using samples at $t = 36$ and $t = 65$ days post-infection. In contrast, we consider escape during $t = 15$ to $t = 36$ days post-infection, with sequence data available only at $t = 36$. Despite the significant differences in method and data, their estimates and ours are similar, with our escape rates ranging from 0.15 to 0.35, depending on the variant.

Our model and hypothesis tests allow us to investigate escape rates across the epitope mutations. The striking feature

Table 6 CH58 frequency data

Day	F	$M1$	$M12$
9	0.71	0.29	0
45	0.11	0	0.89

of Table 4 is the high frequency of H variants with respect to the other epitope mutations. This deviation is reflected in the estimates given in Figure 8, showing escape rates for $H \sim 0.1$ higher than other variants. The higher escape rate of H variants may result from differences in fitness, mutation rate, pMHC binding, or CTL recognition. But before turning to such explanations, it is valuable to consider whether the deviation could be caused by stochastic effects alone.

To address this issue, we can use our null hypothesis methods. Table 5 describes the death rates assumed by a null hypotheses in which all variants have an identical death rate, $0.8 - \delta_M$, during the CTL response time, $[15, 36]$. The parameter δ_M is the escape rate of the variants since $\delta_{F,2} = 0.8$. Aside from this change, all other parameters are the same as above and we still use the escape graph in Figure 1. We considered a range of δ_M from 0 to 0.5; none produced a P -value $> 10^{-4}$, meaning that the data do not fit this null model and accompanying hypothesis that all variants share the same escape rate.

Having rejected the null hypothesis that all mutant variants are identical, a natural second hypothesis would be to somehow group the mutants into several categories. Figure 8 suggests that three groups might be appropriate: (1) H , (2) Q, R, M, C , and (3) N, E, I . The hypothesis can be specified by adapting Table 5 to allow for three different values of the escape rate, δ_M , corresponding to the three groups. We find a P -value of 0.2 when the escape rate is 0.32; 0.24 and 0.16 for H ; Q, R, M, C ; and N, E, I , respectively, suggesting that this particular model is a reasonable fit for the data.

The hypotheses presented above suggest, although certainly do not prove, that CH40 data reflect an escape involving multiple variants with different escape rates. Analysis of this type of escape is not possible under existing methods that assume a single mutant type.

Patient CH58

The data we used for patient CH58 were presented in Goonetilleke *et al.* (2009). We refer the reader to that article and references therein for full details. Briefly, patient CH58 was identified during acute infection approximately 1 week prior to peak viral load. As for the CH40 data set, day 0 represents the time of identification. Blood samples were collected at day 0, day 9, day 45, and subsequent times. Through single-genome amplification, two regions on the founder genome were found to experience early escape: ENV EL9 and ENV 830 (see CH58.e and CH58.g in Figure 2 of Goonetilleke *et al.* 2009). Escape at ENV EL9 was first seen at day 9, at which time ENV 830 and all other regions of escape were homogeneous. By day 45, escape had also

Figure 9 Posterior and deterministic estimates for the escape rates of the CH58 data set. Each tic labeled by an escape rate and its right neighboring tic represent the posterior and deterministic escape-rate estimates, respectively.

occurred at ENV 830. Two other epitopes had low levels of variation by day 45, reflecting the start of escapes that continued past day 45. T-cell assays verified ENV EL9 as a target of T-cell response, but no response to ENV 830 could be found and the nucleotide sequence in that region is not a known CTL epitope, raising the possibility that ENV 830 is the location of compensatory mutations.

At day 9 and day 45, the Goonetilleke *et al.* (2009) data are composed of 7 and 9 sequences, respectively. Importantly, ENV EL9 and ENV 830 are linked in these sequences. Other larger data sets that we are aware of lack linkage information, involve three or more overlapping escapes, or do not capture early stages of escape at the initially targeted epitope. Consequently, despite the small sample size, CH58 data are a useful setting from which to explore the impact of multi-epitope response on escape-rate inference.

We categorized the sequences by whether the ENV EL9 and ENV 830 regions are mutated or of founder type, meaning that we ignored the differences between epitope mutations within each region. (See [File S1](#) for a summary of the different mutations seen at each region.) This gave three variants: founder variants, variants with a mutation in ENV EL9, and variants with a mutation at both ENV EL9 and ENV 830. We represented these variants as F , $M1$, and $M12$, respectively. Note that mutations at only ENV 830, which would be labeled $M2$, were not present in the data and were not included in our model. Table 6 is the reduction of CH58 data to our three variants.

We assumed a double-escape model with the escape graph given in Figure 2. Recall that double-escape models involve three time intervals, $[0, t_A]$, $[t_A, t_D]$, and $[t_D, t'_D]$, where t_A is the time of initial CTL response and t_D, t'_D are the sampling times. We took $t_A = 15$ in units of days post-infection. We assumed that day 0 post-identification corresponds to 20 days post-infection, which gives $t_D = 29$ and $t'_D = 65$ in units of days post-infection. Standard model parameters were chosen to fit the viral load pattern of patient CH58, see Goonetilleke *et al.* (2009, Figure S1), and to fall within known ranges (Perelson *et al.* 1993; Stafford *et al.* 2000; De Boer 2007). Specifically, $k = 2.6 \times 10^{-3}$, $d = 0.01$, $\lambda = 10^8$. We set $\delta_{v,1} = 0.4$, the death rate prior to CTL response, for all variants and $\delta_{M12,2} = 0.4$, the death rate of $M12$ during the time period $[t_A, t_D]$. The escape rates

$\delta_{F,2} - \delta_{M1,2}$, $\delta_{F,3} - \delta_{M1,3}$, and $\delta_{F,3} - \delta_{M12,2}$ were objects of inference. Note that in contrast to CH40, here $M1$ has two escape rates corresponding to the intervals $[t_A, t_D]$ and $[t_D, t'_D]$ and $M12$ has one escape rate corresponding to $[t_D, t'_D]$. As mentioned, since $M12$ variants are not present in the data at day 9 post-identification, we cannot identify the escape rate of $M12$ during both time intervals of CTL response, and so we set the escape rate during $[t_A, t_D]$ to reflect a biologically reasonable value, an assumption that becomes part of the null hypothesis.

Figure 9 gives the posterior and deterministic estimators of the escape rates. As in the case of CH40, we use the posterior estimates to construct hypothesis tests. Due to the small sample size, we consider two hypotheses, labeled A and B . Importantly, we make the true frequencies of F , $M1$, $M12$ variants part of the null hypotheses, meaning that the assumed frequencies are free of sampling error. Table 7 gives the variant frequencies assumed under A and B . In hypothesis A , $M1$ variants at day 45 have frequency 0, while in hypothesis B the $M1$ frequency at day 45 is raised to 0.33 and, correspondingly, the $M12$ frequency lowered to 0.56. Hypotheses A and B represent the endpoints of the 95% confidence interval for $M1$ frequencies at day 45. Table 8 gives the escape rates, chosen using guidance from the posterior and assumed under hypotheses A and B . Hypotheses A and B generated P -values of 0.96 and 0.45, respectively.

For the CH58 data set, the value of our hypothesis tests is not in the P -values generated. Indeed, we chose the escape rates to make the P -values large. Rather, here we have found two models that we believe span a range of possible underlying variant frequencies from which the CH58 data set was sampled.

Figure 10, A and B, shows the dynamics of F , $M1$, $M12$ variant frequencies under hypotheses A and B , respectively. In Figure 10 time is given in days post-identification. At days 9 and 45, the figures show that the variant frequencies are those given in Table 7. Both figures show sudden changes in dynamics at day 9, reflecting the model assumption of changes in death rates between the intervals $[t_A, t_D]$, $[t_D, t'_D]$, where here t_D corresponds to day 9 post-identification.

In both figures, $M1$ variants initially push out F variants, followed by $M12$ variants pushing out both $M1$ and F variants. To frame the dynamics intuitively, we can suppose

Table 7 Frequencies under hypotheses A and B

Hypothesis	Day post-identification	<i>F</i>	<i>M1</i>	<i>M12</i>
<i>A</i>	9	0.71	0.29	0
	45	0.11	0	0.89
<i>B</i>	9	0.71	0.29	0
	45	0.11	0.33	0.56

that CTLs initially respond to the ENV EL9 epitope, therefore selecting for *M1* and *M12* variants. This response lasts until day 9 and then is replaced by a secondary response that selects for *M12* variants over *M1* variants. In the case of hypothesis *A* dynamics, the secondary response selects *F* variants over *M1* variants, perhaps meaning that CTL response completely shifted to a new epitope after day 9. In contrast, for hypothesis *B* dynamics, the secondary response selects *M1* variants over *F* variants, corresponding intuitively to a weakened, but still present, CTL response to ENV EL9. The figures also provide a feel for the dynamics that would have resulted from *M1* frequencies between those of hypotheses *A* and *B*.

The dark curve in both figures corresponds to the frequency of mutants at ENV EL9. The curves both have values of 0.29 and 0.89 at days 9 and 45 post-identification, exactly the value of the CH58 data set. If we restricted our attention only to ENV EL9, meaning that *M1* and *M12* variants would be indistinguishable, the dark curves are the profiles we would see. In this situation, using existing methods (Asquith *et al.* 2006), we find an escape rate of 0.08 (similar to the value found in Ganusov *et al.* 2011 and Gonnelleke *et al.* 2009), which is significantly less than the first *M1* escape rate and the *M12* escape rate assumed in the case of null hypotheses *A* and *B*.

The dynamics of Figure 10 show that between day 9 and 45 post-infection, ENV EL9 escape is driven by ENV 830 selection. Indeed, under both hypothesis, *M1* variants eventually drop in frequency while *M12* variants escape at a rate of 0.24, suggesting strong selection for ENV 830 mutations

Table 8 Escape rates under hypotheses A and B

Hypothesis	First <i>M1</i> escape rate	Second <i>M1</i> escape rate	<i>M12</i> escape rate
<i>A</i>	0.46	−0.1	0.24
<i>B</i>	0.46	0.06	0.24

and weak or negative selection for ENV EL9 mutations. In this scenario, the escape rate estimated by the ENV EL9 data alone is influenced more by secondary responses not directly associated to ENV EL9 than by the ENV EL9 CTL response itself.

Discussion

We have investigated two ways in which current escape-rate inference methods can be extended. First, through the CH40 data set, we considered escape through multiple mutations on a single epitope. We found that escape rates varied between the different epitope mutations. More specifically, our posterior construction gave different escape rates for the different epitope mutations and our hypothesis tests rejected a null model under which all mutations shared the same escape rate. Current techniques do not allow for such analysis, since all epitope mutations must be grouped into a single mutant type. Second, through the CH58 data set, we considered escape in response to selection at two genome regions (one an epitope, the other region possibly not). In this case, we found that escape dynamics at the two regions were intertwined, raising the possibility that existing methods applied to a single epitope may be biased due to escape at other epitopes.

While we have presented a basic framework for computational inference of HIV escapes, two current limitations stand out. First, the computations we have presented consider only one epitope with many escape mutations or two epitopes with one escape mutation each. One would like to consider the realistic case of escape at multiple epitopes through multiple mutations. While our methods should apply to this greater context, understanding how to construct escape graphs and

Figure 10 Frequency dynamics of *F*, *M1*, *M12* variants under (A) hypothesis *A* and (B) hypothesis *B*. The thick line is the frequency of ENV EL9 mutants and represents the data seen if only the ENV EL9 epitope is considered.

birth–death processes that can extract biologically useful information in this more computationally challenging setting requires further work. Second, as has been noted above, our inference methods currently require data sets with a large number of sampled sequences. For escape at a single epitope, deep-sequencing data sets provide the needed sample sizes and our methods can be applied. However, for escape at multiple epitopes, deep-sequencing data sets usually fail to provide linkage information. In our analysis of the CH58 data set, this restriction forced us to include more information in our null hypothesis. In principle, extension to small sample size should be possible, but more theoretical and computational work is needed.

Acknowledgments

I thank Mark Maloof and Jami Montgomery for innumerable discussions regarding the C++ implementation of the algorithms discussed. I thank Ali Arab for several conversations that helped me clarify the hypothesis testing approach taken. I thank Jessica Conway for discussions and advice relating to the Fourier inversion methods used to generate the pop size distributions. This work was supported by National Science Foundation grant DMS-1225601.

Literature Cited

- Asquith, B., C. Edwards, M. Lipsitch, and A. McLean, 2006 Inefficient CTL mediated killing of HIV-1 infected cells in-vivo. *PLoS Biol.* 4: 583–592.
- Batorsky, R., M. Kearney, S. Palmer, F. Maldarelli, I. Rouzine *et al.*, 2011 Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. *Proc. Natl. Acad. Sci. USA* 108: 5661–5666.
- B. Bimber, B. Burwitz, S. O'Conner, A. Detmer, and E. Gostick *et al.*, 2009 Ultradeep pyrosequencing detects complex patterns of CD8 T-lymphocyte escape in simian immunodeficiency virus-infected macaques. *J. Virol.* 83: 8247–8253.
- C. L. Boutwell, M. M. Rolland, J. T. Herbeck, J. L. Mullins, and T. M. Allen, 2010 Viral evolution and escape during acute HIV-1 infection. *J. Infect. Dis.* 202: 309–314.
- Cohen, M., G. Shaw, A. McMichael, and B. Haynes, 2011 Acute HIV-1 infection. *N. Engl. J. Med.* 364: 1943–1954.
- Conway, J., and D. Coombs, 2011 A stochastic model in latently infected cell reactivation and viral blip generation in treated hiv patients. *PLOS Comp. Bio.* 7: 1–15.
- De Boer, R., 2007 Understanding the failure of CD8 T-cell vaccination against simian/human immunodeficiency virus. *J. Virol.* 81: 2838–2848.
- Desai, M., and D. Fisher, 2007 Beneficial mutation-selection balance and the effect of linkage on positive selection. *Genetics* 176: 1759–1798.
- Fernandez, C., I. Stratov, R. De Rose, K. Walsh, C. Dale *et al.*, 2005 Rapid viral escape at an immunodominant simian-human immunodeficiency CTL epitope exacts a dramatic fitness cost. *J. Virol.* 79: 5721–5731.
- Fiebig, E., D. Wright, B. Rawal, P. Garrett, R. Schumacher *et al.*, 2003 Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. *AIDS* 17: 1871–1879.
- Fisher, W., V. Ganusov, E. Giorgi, P. Hraber, B. Keele *et al.*, 2010 Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE* 5: 1–15.
- Frost, S. M., R. Nijhuis, R. Schuurman, C. Boucher, and A. L. Brown, 2000 Evolution of lamivudine resistance in HIV-1 infected individuals: the relative roles of drift and selection. *J. Virol.* 74: 6262–6268.
- Ganusov, V., and R. De Boer, 2006 Estimating costs and benefits of CTL escape mutations in SIV/HIV infection. *PLOS Comput. Biol.* 2: 182–187.
- Ganusov, V., N. Goonetilleke, M. Liu, G. Ferrari, G. Shaw *et al.*, 2011 Fitness costs and diversity of the CTL response determine the rate of CTL escape during acute and chronic phases of HIV infection. *J. Virol.* 85: 10518–10528.
- Gilks, W., S. Richardson, and D. Spiegelhalter, 1996 *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.
- Gillespie, D., 2001 Approximate accelerated simulation of chemically reaction systems. *J. Chem. Phys.* 81: 1716–1733.
- Gillespie, J., 1991 *The Causes of Molecular Evolution*. Oxford University Press, Oxford, UK.
- Goonetilleke, N., M. Liu, J. Salazar-Gonzalez, G. Ferrari, E. Giorgi *et al.*, 2009 The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *J. Exp. Med.* 206: 1253–1272.
- Goulder, P., and D. Watkins, 2004 HIV and SIV CTL escape: implications for vaccine design. *Nat. Rev. Immunol.* 4: 630–640.
- Henn, M. R., C. L. Boutwell, P. Charlebois, N. J. Lennon, K. A. Power, *et al.*, 2012 Whole genome sequencing of HIV-1 reveals impact of early minor immune variants on immune recognition during acute infection. *PLoS Pathog.* 8: 1–14.
- Kaplan, N., T. Darden, and R. Hudson, 1988 The coalescent process in models with selection. *Genetics* 120: 819–829.
- Keele, B., E. Giorgi, J. Salazar-Gonzalez, J. Decker, K. Pham *et al.*, 2008 Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc. Natl. Acad. Sci. USA* 105: 7552–7557.
- Krone, S., and C. Neuhauser, 1997 Ancestral processes with selection. *Theor. Popul. Biol.* 51: 210–237.
- Lemey, P., M. Salemi, and A.-M. Vandamme, 2009 *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, Ed. 2. Cambridge University Press, Cambridge, UK.
- Leviyang, S., 2012 Sampling HIV intrahost genealogies based on a model of acute stage CTL response. *Bull. Math. Biol.* 3: 509–535.
- Liu, Y. *et al.*, 2006 Selection on the human immunodeficiency virus type 1 proteome following primary infection. *J. Virol.* 80: 9519–9529.
- Mansky, L., 1996 Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. *AIDS Res. Hum. Retroviruses* 12: 307–314.
- Mehandru, S., M. Poles, K. Tenner-Racz, A. Horowitz, A. Hurley *et al.*, 2004 Primary HIV-1 infection is associated with preferential depletion of CD4+ T lymphocytes from effector sites in the gastrointestinal tract. *J. Exp. Med.* 200: 761–770.
- Mehandru, S., M. Poles, and K. Tenner-Racz, V. Manuellil, P. Jean-Pierre *et al.*, 2007 Mechanisms of gastrointestinal CD4+ T-cell depletion during acute and early human immunodeficiency virus type 1 infection. *J. Virol.* 81: 599–612.
- Merrill, S., 2005 The stochastic dance of early HIV infection. *J. App. Comp. Math.* 184: 242–257.
- Miao, H., X. Xia, A. Perelson, and H. Wu, 2011 On identifiability of nonlinear ODE models with application in viral dynamics. *SIAM Rev.* 53: 3–39.
- Nielsen, R., and Z. Yang, 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148: 929–936.

- Nowak, M., 2000 *Evolutionary Dynamics: Exploring the Equations of Life*. Oxford University Press, Oxford.
- Nowak, M., and R. May, 2000 *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford University Press, Oxford.
- Pennings, P., and J. Hermisson, 2006 Soft sweeps iii: the signature of positive selection from recurrent mutation. *PLoS Genet.* 2: 1998–2012.
- Perelson, A., 2002 Modeling viral and immune system dynamics. *Natl. Rev.* 2: 28–36.
- Perelson, A., D. Kirschner, and R. De Boer, 1993 Dynamics of HIV infection of CD4+ T cells. *Math. Biosci.* 114: 81–125.
- Perelson, A., A. Neumann, M. Markowitz, J. Leonard, and D. Ho, 1996 HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* 271: 1582–1586.
- Ribeiro, R., and S. Bonhoeffer, 1999 A stochastic model for primary HIV infection: optimal timing of therapy. *AIDS* 13: 351–357.
- Rouzine, I., and J. Coffin, 2010 Many-site adaptation in the presence of infrequent recombination. *Theor. Popul. Biol.* 77: 189–204.
- Rouzine, I., A. Rodrigo, and J. Coffin, 2001 Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. *Microbiol. Mol. Biol. Rev.* 65: 151–185.
- Slatkin, M., and R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129: 555–562.
- Stafford, M., L. Corey, Y. Cao, E. Daar, D. Ho *et al.*, 2000 Modeling plasma virus concentration during primary HIV infection. *J. Theor. Biol.* 203: 285–301.
- Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.
- Tuckwell, H., P. Shipman, and A. Perelson, 2008 The probability of HIV infection in a new host and its reduction with microbicides. *Math. Biosci.* 214: 81–86.
- Wick, D. and S. G. Self, 2000 Early HIV infection in vivo: branching-process model for studying timing of immune responses and drug therapy. *Math. Biosci.* 165: 115–134.
- Zhuo, H., and K. Dorman, 2005 A branching process model of drug resistant HIV, pp. 457–496 in *Deterministic and Stochastic Models of AIDS Epidemics and HIV infections with Intervention*, Chap. 18 edited by T. Wai-Yuan, and H. Wu. World Scientific Publishing Co., Singapore.

Communicating editor: R. Nielsen

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.150862/-/DC1>

Computational Inference Methods for Selective Sweeps Arising in Acute HIV Infection

Sivan Leviyang

File S1

Supporting Information

S1. Identifiability

For deterministic dynamical systems, identifiability roughly means that parameters can be uniquely determined from some portion of the system dynamics, see [6] for a more precise definition. Connecting this idea to typical HIV datasets, identifiability means that given time points at which data is collected, parameters can be uniquely determined if the system state at those time points is known. Importantly, identifiability does not mean that parameters can be uniquely identified from actual finite datasets which include sampling variance.

For our model, identifiability is more complex because stochasticity is an essential part of the dynamics rather than just the dataset. However, our stochastic reduction simplifies the stochasticity to a function of the pop size samples and we can interpret identifiability to mean that parameters are uniquely determined if the the system state at sample time points is known and, additionally, if the pop size sample values are known.

We now show that identifiability of all the death rate parameters δ_v is not possible. Since we can assume that our pop sizes are known, our model reduces to a deterministic model almost identical to the standard model. As a simple example then, consider the standard model (1) with two variants, F and M . Our data is in the form of frequency information, so we set $f_F(t) = I_F(t)/(I_F(t) + I_M(t))$ and $f_M(t) = I_M(t)/(I_F(t) + I_M(t))$ where $f_F(t)$ and $f_M(t)$ are the fraction of F and M variants, respectively, at time t . Matching our assumptions for the single escape model, consider the standard model on $[t_A, t_D]$ and assume that the system state is known at those two time points. We can solve the ODEs of (1) to find

$$I_M(t_D) = I_M(t_A) \exp[H - \delta_M(t_D - t_A)]. \quad (\text{S1})$$

where $H = \int_{t_A}^{t_D} kT(t)dt$. Simple algebra then gives,

$$f_F(t_D) = \frac{1}{1 + \left(\frac{1}{f_F(t_A)} - 1\right) \exp[(\delta_F - \delta_M)(t_D - t_A)]}, \quad (\text{S2})$$

meaning that frequency data uniquely determines $\delta_F - \delta_M$, not both δ_F and δ_M .

The same argument applies to our model, with the exception that H needs to be redefined to include mutation, i.e. (2) must be solved rather than the I_v equation of (1). If mutation rates do not differ between variants, the same argument as above holds for our model. If mutation rates do differ between variants, the single H must be replaced by an H_v for each v . This may allow for identifiability, but we have found through numerical experiments that the stochasticity of the pop size samples and datasets is far greater than the differences in the H_v , thereby making practical identifiability (see [6] for a discussion of practical identifiability) of all δ_v difficult. For these reasons, in all cases we take a cautious approach and simply infer the differences $\delta_F - \delta_v$.

S2. Errors of Stochastic-Deterministic Decomposition and Stochastic Reduction

Previous authors have considered algorithms similar to ours, except that the pre-stochastic and stochastic interval steps are replaced by Gillespie simulation [10, 7, 3, 5]. In that setting, for

selection strengths seen during HIV escape, relative error for the deterministic step has been shown to be approximately .03 when $L = 100$ [5, 3, 10].

To assess error more precisely in our setting, we performed some numerical experiments using the escape graph of Figure 2. Figure S1A shows the probability density of $I_{M1}(15)$, the number of $M1$ variant infected cells at $t = 15$, computed using the stochastic-deterministic decomposition with $L = 2000$, $\epsilon = 0$. Parameters were set at $k = 2 * 10^{-3}$, $d = .01$, $\lambda = 10^8$, and $\delta_v = .4$ for $v = F, M1, M12$. These parameter choices fall within the typical range used for HIV, e.g. [11, 8, 2, 1], with the caveat that λ, k must be converted to per body units and remembering that we assume virions are in steady state. (For example, the $k = 2 * 10^{-3}$ in our model corresponds to $k = 1.2 * 10^{-5}$ in the model of [1], approximately the value cited in that paper.) The stochastic interval for $M1$ was [4.6, 13.7], so by $t = 15$ the $M1$ variants were in the deterministic portion of the decomposition. Figure S1B shows the difference in the cumulative distribution between using $L = 2000$ and an exact Gillespie distribution run up to $t = 15$. As can be seen the maximum difference in the cdfs is less than .01. In the Results section, we make the conservative choice of $L = 2000$, although $L = 100$ is likely sufficient.

Figure S1: **Accuracy of Stochastic-Deterministic Decomposition.** (A) Distribution of $I_{M1}(15)$, the number of $M1$ variants at $t = 15$, with parameters $k = 2.6 * 10^{-3}$, $d = .01$, $\lambda = 10^8$ and $\delta_v = .4$ for $v = F, M1, M12$. (B) Difference of cumulative distributions for $I_{M1}(15)$ using Gillespie simulation and stochastic-deterministic decomposition with $L = 2000$. Both (A) and (B) were generated using 10^6 simulations.

S3. Parameters for Figure 4

Figure 4 assumed only two attack intervals, $[0, 15]$, $[15, 30]$. We ran the model up to time 30, although the full dynamics are not shown since the M12 stochastic interval always ended before time 20. In interval $[0, 15]$, we assumed death rates of .4 for all variants. In interval $[15, 30]$ we assumed death rates of 1.1, .7, .4 for F , $M1$, $M12$, respectively. We set $k = 2.8 * 10^{-3}$, $d = .01$, $\lambda = 10^8$. We set $L = 1000$, $\epsilon = .03$.

S4. Computing Overall p-Value from Multiple q 's

To describe how we combine different q 's to form an overall p -value, consider the case where for a given hypothesis θ we have two q 's, q_1, q_2 , both uniformly distributed on $[0, 1]$. Let Ω be the unit square produced by combining all pairs of numbers both falling in $[0, 1]$. Let $f(x)$ for $x \in [0, 1]$ be the shortest distance from x to either 0 or 1, i.e. $f(x) = \min(x, 1 - x)$. Then we define the p -value associated with θ by,

$$\text{p-value of } \theta = P(f(u_1)f(u_2) < f(q_1)f(q_2)) \quad (\text{S3})$$

where u_1, u_2 are uniform samples from $[0, 1]$. In words, $f(q_1)f(q_2)$ is a measure of how far the pair is from the boundary of Ω and we set the p -value to be the probability that two uniform samples from $[0, 1]$ are closer to the boundary. For example, the pair .5, .5 will have a measure of $.5^2$ from the boundary of Ω which is greater than any other point. Hence .5, .5 would have a p -value of 1. In practice, we evaluate (S3) through Monte Carlo methods by sampling 10^7 pairs of u_1, u_2 and determining the fraction of times $f(u_1)f(u_2) < f(q_1)f(q_2)$. For escape graphs with more variants, the same idea is employed but in higher dimension.

S5. Numerical Experiments - Parameter Inference

The numerical experiment used to examine parameter inference error consisted of the following steps:

1. We chose values for model parameters in a biologically reasonable range. Specifically k, λ, d were chosen independently and uniformly from $[10^{-3}, 5 * 10^{-3}]$, $[10^8, 10^9]$, $[.01, .1]$, respectively [11, 2, 8]. Each $\delta_{v,1}$ (the death rate of variant v prior to CTL response) was chosen uniformly from $[.4, .5]$ [9]. t_A , the CTL response time, was sampled uniformly from $[13, 17]$. We chose $\delta_{F,2}$ uniformly on $[.6, 1.2]$ and $\delta_{v,2}$ for each $v \neq F$ in the escape graph uniformly and independently from $[.4, \delta_{F,2}]$.
2. Using the parameter values chosen, we simulated our model and produce a simulated dataset. The simulated datasets consisted of F and v variant frequencies at time $t_D = 40$. In forming the simulated dataset, we assumed 5000 sequences were sampled at time t_D , a reasonable number for deep sequencing datasets.
3. We then used our simulated dataset to construct a posterior for the escape rates assuming no knowledge of our samples for $\delta_{F,2}, \delta_{v,2}$ nor the distributions from which the samples were drawn. The posterior was constructed by running an MCMC algorithm 10^5 time steps.

To compare our posterior method with a deterministic method, we also performed an experiment using steps 1 – 2 as above, but with step 3 replaced by

- Given the simulated dataset constructed in step 2, we used the standard model (1) with the I_v equation replaced by (2) to compute a least-squares estimate of escape rates. Stochasticity corresponding to sampling was preserved. Using the same approach as [4], we produced a distribution of estimates by resampling sequences according to the variant frequencies of the simulated dataset. Each resample produced a new dataset and a corresponding least squares estimate. We resampled 10000 times to form a distribution of estimates.

We repeated the above experiments 1000 times for both posterior and deterministic methods. Generating 1000 posterior experiments took about a day, in the case of 1 epitope mutation, and about 3 days, in the case of 8 epitope mutations. The 1000 experiments based on the deterministic method took under 3 hours in all cases.

Our distributions are multidimensional, e.g. in the case of 3 epitope mutations we have a 3 dimensional distribution. To form the relative error, we formed the relative error for the marginal distributions of the escape rate, $\delta_{F,2} - \delta_{v,2}$, for each variant v . We collected these marginal error distribution over all variants v and all experiments to produce the figure 6.

To generate Figure 7, we used a 3 epitope mutation, single escape model with $k = 4 * 10^{-3}$, $d = .01$, $\lambda = 10^8$, $\delta_{v,1} = .4$ for all variants, $\delta_{F,2} = .76$ and the death rates of the three epitope mutations during CTL response given by .65, .49, .58. $t_A = 15$, $t_D = 40$. We ran this model generating at time t_D the variant frequencies given in Table S1.

Table S1: Dataset of Figure 7

F frequency	mutation variant frequencies		
.63	.01	.32	.04

Double Escape Experiments:

We considered the escape graph of Figure 2 with the goal of estimating the 1st and 2nd $M1$ escape rates and the $M12$ escape rate which are given by $\delta_{F,2} - \delta_{M1,2}$, $\delta_{F,3} - \delta_{M1,3}$, and $\delta_{F,3} - \delta_{M12,3}$, respectively. Recall that $\delta_{v,2}$ represents the death rate of variant v during the first time interval of CTL response $[t_A, t_{D1}]$, while $\delta_{v,3}$ represents the death rate between the first and second sampling times, $[t_{D1}, t_{D2}]$. As mentioned, $\delta_{M12,2}$ and $\delta_{M12,3}$ are not both identifiable, so we set $\delta_{M12,2}$ to a specific value and infer only a single escape rate for $M12$.

We ran essentially the same experiments as we did for the single escape case. Specifically, We chose k, d, λ, t_A as in the numerical experiments for single escape graphs. $\delta_{v,1}$ was chosen uniformly from $[.4, .45]$ independently for all v . We chose a number a_1 , uniformly from $[.2, .6]$. This represented the death rate due to CTL attack at epitope 1. Then $\delta_{F,2} = .4 + a_1$, and $\delta_{M1,2}, \delta_{M12,2}$ are selected uniformly from $[.4, .4 + a_1]$. $M1$ and $M12$ evade the attack, but perhaps not completely. We then chose a_2 uniformly from $[.2, .5]$, representing the strength of attack at a second epitope. Then $\delta_{F,3} = .4 + a_1 + a_2$, $\delta_{M1,3} = \delta_{M1,2} + a_2$, $\delta_{M12,3} = \delta_{M12,2} + u$ where u is uniform on $[0, a_2]$. t_D, t'_D were set to 30, 50, respectively.

Figures S2A-C gives a box plot for the relative errors of the three escape rates averaged over 1000 experiments. Note that the y-axes in subfigure *C* is on a much larger scale than those of subfigures *A* and *B*. In each of the figures, the left and right box plots are generated using our

posterior and the deterministic method, respectively. We found that the escape rate of $M12$ is biased down under both estimators. This is because overestimation of the 1st $M1$ escape rate produced a large underestimation of the $M12$ escape rate, while underestimation of the 1st $M1$ escape rate produced only a modest overestimation of the $M12$ escape rate. In about half the cases, when the 1st $M1$ escape rate was unbiased or underestimated, the $M12$ escape rate was roughly unbiased (data not shown).

Figure S2: **Relative error for escape rate estimates under the double escape model of Figure 2.** For each escape rate, the left(right) boxplot gives the absolute error scaled by .4 produced by the posterior(deterministic) estimation method. The figure is an average of 1000 experiments. Note the difference in the y-axis scale in subfigure *C*

Figure S3 gives the posterior and deterministic estimates for a single simulated dataset, rather than the averaged results of Figure S2. The general biases are the same in both figures, but both the posterior and the deterministic estimator distributions are much narrower for a single dataset. Note that the correct value, represented here as an error of 0, is within all the posterior distributions, but that is not the case for the deterministic distributions.

S6. Numerical Experiments - Hypothesis Testing

To examine our hypothesis testing methods under the double escape graph of Figure 2, we ran type I error and power experiments as we did for the single escape model. Table S2 gives the type I error under different sample sizes and confidence levels. The results are generally similar to the single escape model, with an ∞ sample size giving correct type I errors, while sample sizes of 5000 and 1000 are too conservative in accepting the null hypothesis. More specifically type I errors given in Table S2 are similar to those under the single escape model with 3 epitope mutations. As in the case of a single escape model, type I errors go to zero as confidence levels go to 1. For example, for a sample size of 5000, a confidence level of .999 produced a type I error of .01. Power as a

Figure S3: **Escape rate estimates for a single simulated dataset under the double escape model of Figure 2.** Note the difference in the y-axis scale in subfigure *C*

function of the euclidean distance between the null and alternative model death rates, i.e. the $\delta_{v,2}$'s and $\delta_{v,3}$'s, is given in Table S3 and is similar to the single escape model case.

S7. CH58 Epitope and Mutations

Table S4 gives the CH58 dataset at ENV EL9 and ENV 830 for days 9 and 45. The two epitopes are linked in the dataset.

S8. Computing the Pop Size Distribution

Computing pop size distributions reduces to computing the distribution of a birth-death process with immigration, see [12] for discussion of such processes. Here, the role of mutation in forming the pop size distribution is played by immigration. Let $b(t)$, $d(t)$, $m(t)$ be the birth, death, and immigration rates where $b(t)$ is the rate at which an individual produces a single offspring, $d(t)$ is the rate at which an individual dies, and $m(t)$ is the rate at which an immigrant joins the population. Let $B(t)$ be the associated process and set $B(0) = 0$.

The characteristic function for $B(t)$ is

$$E[\exp[i\omega B(T)]] = \exp\left[-\int_0^T ds m(s)(1 - E[\exp[i\omega K(s, T)])]\right] \quad (S4)$$

where, to define $K(s, T)$, let $Y(t)$ be a birth-death process with birth, death rates $b(t)$, $d(t)$ respectively and no immigration. Let $Y(s) = 1$, then $K(s, T) = Y(T)$.

Table S2: Type I Error Estimates : Double Escape Model

sample size	confidence level	
∞	99%	.02
	95%	.06
	90%	.11
	80%	.20
	50%	.45
5000	99%	.05
	95%	.15
	90%	.21
	80%	.30
	50%	.55
1000	99%	.09
	95%	.28
	90%	.36
	80%	.45
	50%	.67

Table S3: Power Estimates : Double Escape Model

distance to null model						
.01-.03	.03-.06	.06-.09	.09-.12	.12-.15	0.15-.18	.18-.21
.52	.6	.75	.96	.98	1	1

To compute the density of $B(t)$, we invert the characteristic function. Applying such an inversion gives the following,

$$\begin{aligned}
 P(B(T) = x) &= \frac{1}{2\pi} \int_0^{2\pi} d\omega \exp[-i\omega x] \exp[i\omega B(T)] \\
 &= \frac{1}{2\pi} \int_0^{2\pi} d\omega \exp[-i\omega x] \exp\left[-\int_0^T ds m(s) (1 - E[\exp[i\omega K(s, T)])]\right] \\
 &= \frac{1}{\pi} \text{Real} \left(\int_0^{\pi} d\omega \exp[-i\omega x] \exp\left[-\int_0^T ds m(s) (1 - E[\exp[i\omega K(s, T)])]\right) \right),
 \end{aligned} \tag{S5}$$

where, as observed in [1], the last relation follows by noting that values of the integrand for $\omega \in [0, \pi]$ are conjugate to those for $\omega \in [\pi, 2\pi]$.

$K(s, T)$ is described by a binary branching process, so its characteristic function can be expressed as the solution to a Kolmogorov equation. Let $H(s, T) = E[\exp[i\omega K(s, t)]]$, then

$$\frac{d}{ds} H(s, T) = -(b(s) + d(s))H(s, T) + b(s)H^2(s, T) + d(s) \tag{S6}$$

with $H(T, T) = \exp[i\omega]$. This equation can be solved using a standard ODE solver. In turn, (S5)

Table S4: CH58 Sequence Data

day	a.a. sequence	number of samples
9	ERYLRDQQL...RFIRAVLH	5
9	—G—...—	1
9	—H—...—	1
45	—...—	1
45	—H—...—T—	1
45	—H—...—I—	1
45	—H—...—V—	1
45	—K—...—T—	1
45	—S—...—V—	1
45	—S—...—V—	1
45	—G—...—I—	1
45	—G—...—M—	1

can be computed using numerical integration. We perform both in C++ using the Gnu Scientific Library.

A slightly faster approach is to scale $B(T)$ by L , so that we actually compute the density of $B(T)/L$. Numerically this is better because we know $B(T)$ is $O(L)$ and so the density of $B(T)/L$ is roughly unchanged as L grows. This allows us to optimize the grid on which we perform the numerical integration. Put another way, there is usually no reason to integrate (S5) in ω all the way to π because $E[\exp[i\omega B(T)]]$ quickly collapses and since the value $\int_0^T ds m(s)$ is large, the integrand collapses as well. Roughly we can stop when $\omega(L) \gg 1$, meaning that we can integrate in ω to $O(1/L)$. When we consider $B(T)/L$ and change the integrand correspondingly, we need to integrate in ω to $O(1)$ roughly independent of L . In practice we integrate to 80.

References

- [1] J. CONWAY AND D. COOMBS, *A stochastic model in latently infected cell reactivation and viral blip generation in treated hiv patients*, PLOS Comp. Bio., 7 (2011).
- [2] R. DE BOER, *Understanding the failure of cd8 t-cell vaccination against simian/human immunodeficiency virus*, J. Virology, 81 (2007), pp. 2838–2848.
- [3] M. DESAI AND D. FISHER, *Beneficial mutation-selection balance and the effect of linkage on positive selection*, Genetics, 176 (2007), pp. 1759–1798.
- [4] V. GANUSOV, N. GOONETILLEKE, M. LIU, G. FERRARI, G. SHAW, ET AL., *Fitness costs and diversity of the ctl response determine the rate of ctl escape during acute and chronic phases of hiv infection*, J. Virology, 85 (2011), pp. 10518–10528.
- [5] S. LEVIYANG, *Sampling hiv intrahost genealogies based on a model of acute stage ctl response*, Bull. Math. Bio., 3 (2012), pp. 509–35.

- [6] H. MIAO, X. XIA, A. PERELSON, AND H. WU, *On identifiability of nonlinear ode models with application in viral dynamics*, SIAM Review, 53 (2011), pp. 3–39.
- [7] M. NOWAK, *Evolutionary Dynamics : Exploring the Equations of Life*, Oxford University Press, 2000.
- [8] A. PERELSON, D. KIRSCHNER, AND R. DE BOER, *Dynamics of hiv infection of cd4+ t cells*, Math. Biosci., 114 (1993), pp. 81–125.
- [9] A. PERELSON, A. NEUMANN, M. MARKOWITZ, J. LEONARD, AND D. HO, *Hiv-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time*, Science, 271 (1996), pp. 1582–1586.
- [10] I. ROUZINE, A. RODRIGO, AND J. COFFIN, *Transition between stochastic evolution and deterministic evolution in the presence of selection : General theory and application to virology*, Microbiology and Molecular Biology Reviews, (2001), pp. 151–185.
- [11] M. STAFFORD, L. COREY, Y. CAO, E. DAAR, D. HO, ET AL., *Modeling plasma virus concentration during primary hiv infection*, J. Theor. Biol., 203 (2000), pp. 285–301.
- [12] S. TAVARE, *Line-of-descent and genealogical processes, and their applications in population genetics models*, Theo. Pop. Bio., 26 (1984), pp. 119–163.