Sampling HIV Intrahost Genealogies Based on a Model of Acute Stage CTL Response

Sivan Leviyang*

Abstract

Cytotoxic T lymphocytes (CTLs) play an important role in the immune response to HIV during the acute stage of infection, but the effect of CTLs on HIV intrahost genetic diversity is poorly understood. We introduce a model of CTL attack during the acute stage. Assuming this model, we develop a method to sample HIV intrahost genealogies. Using our sampling approach we characterize the evolutionary forces that shape HIV genealogies. In particular, we show that early mutation events can have significant impact on HIV genealogies and that certain types of CTL attack are best at controlling HIV genetic diversity. Our sampler represents a first step towards using HIV genetic data to infer properties of CTL attack.

Key Words: HIV, CTL, coalescent.

1 Introduction

Cytotoxic T lymphocytes (CTLs) have been shown to play an important role in shaping and possibly controlling HIV infection during the initial, acute stage of infection. Experimental data shows initial CTL response to be temporally correlated with peak viral load [5, 18], suggesting a role for CTLs in controlling HIV infection. Further, CTLs have been shown to be effective killers of HIV infected cells during the acute stage [10]. In SIV studies, prevention of a CTL response eliminated the drop in viral load often seen in acute HIV infection [33]. And in considering HIV infection generally, HLA association with delayed progression to AIDS suggests an important role for CTL response throughout HIV infection [6]. Yet besides studies that link specific epitopes to CTL attack, e.g. [15], little is known about the impact of CTL attack on intrahost HIV genetic diversity.

Theoretical models of CTL attack have been studied for over twenty years (see [31, 29] for a review). Various models of CTL attack are possible, but typically a dynamical system composed of HIV virions, infected and uninfected CD4 cells, and CTLs targeting infected CD4 cells have been used. Such models

*Georgetown University. Department of Mathematics. sr286@georgetown.edu
are usually deterministic and the dynamical system variables correspond to population sizes, i.e. the number of HIV virions, the number of infected CD4 cells, the number of CTLs targeting a specific HIV epitope, etc. Stochastic models have also been studied, although typically only of the initial stages of HIV infection prior to CTL response, e.g. [25, 37]. HIV infection can be investigated through dynamical system models by fitting the system parameters to experimental viral load data, an approach that has produced some major breakthroughs, e.g. [30].

Over the past decade, HIV genetic data has become nearly as plentiful as viral load data and one would like to use such genetic data to understand CTL attack. Various methods exist to analyze genetic data, with perhaps the three most popular being distance, phylogenetic, and coalescent based methods (see chapters 5, 6 and 17 in [22] for a review of these three approaches). Distance and phylogenetic based methods are used to infer an underlying genealogy or mutation rate, but aside from a mutation model they do not contain an underlying evolutionary model and so are not applicable to analyzing the effect of CTL attack on HIV genetic diversity. Coalescent based methods address such issues by assuming an explicit model for different evolutionary forces acting on a population, examples of such forces include genetic drift, migration, mutation, and selection. For each evolutionary model there will be an associated probability distribution over the set of possible genealogies formed from a given number of samples. Coalescent based methods use this underlying genealogy distribution to analyze the evolutionary forces acting on a population. In the case of a Wright-Fisher population with very large population sizes, the genealogy distribution is well described by a limiting distribution often referred to as Kingman’s coalescent or simply the coalescent [17]. To avoid confusion, hereafter we use the term coalescent to mean any genealogy distribution as opposed to the specific coalescent which we refer to as Kingman’s coalescent.

Different coalescent based methods have been used to analyze different aspects of HIV infection, e.g. [20, 1, 21, 14]. However, current coalescent based methods are not built on models that include the role of CTL attack. As an example, the programs BEAST and LAMARC assume a Kingman coalescent or a closely associated coalescent with varying population size (in the case of BEAST) and migration (in the case of LAMARC) [7, 19]. As mentioned, the Kingman coalescent is associated with Wright-Fisher populations and does not explicitly model CTL attack. In recent work [14], Keele et al. used coalescent based methods to show that initial HIV infection is most often associated with a single infecting genome, however the sampling times for the genetic data came prior to CTL response and therefore the coalescent considered was a variant of the Kingman coalescent that assumes exponential growth, the so-called Slatkin-Hudson coalescent [34].

In this paper we introduce a model of initial CTL attack and then characterize the associated genealogy distribution by describing an algorithm that samples from this distribution. The algorithm is implemented in C++ and code is available upon request. To explain the utility of such a sampler, we distinguish between two types of coalescent based methods. In many coales-
cent based methods, one simply wants to understand how different evolutionary forces affect the genetics of a population. For example, if one wants to understand how specific evolutionary forces affect the heterozygosity of a population, then sampling from the genealogy distribution is all that is required, see [36] for many examples of such investigations. However, often one would like to infer a genealogy or an underlying parameter of the model given a specific set of genetic data. In this setting, current methods depend on determining the probability of a given genealogy rather than sampling an arbitrary genealogy. In this paper we do not explicitly address this important issue, but we note that the same methods used to build our sampler can be used to build an algorithm that determines the probability of a given genealogy.

1.1 Definitions

To specify our model, we start by introducing three time points over which our model is configured: infection time, peak viral load time, and sampling time. We measure time in units of 2 days, the half life of an HIV infected cell [24]. We take infection time to occur at time \( t = 0 \) and we assume infection by a single HIV genome, a common situation as demonstrated in [14]. We refer to the infecting HIV genome as the founder genome. Letting \( T_{\text{peak}} \) be the time of peak viral load we set, in our time units, \( T_{\text{peak}} = 20 \), a value typical of many HIV infections [35]. We assume that CTL response first occurs at time \( T_{\text{peak}} \) and is limited to two possible epitopes, \( e_1 \) and \( e_2 \), both of which are found on the founder genome. We choose attack at two epitopes as an initial case in which to develop our sampler, there is nothing inherent to the model that prevents us from considering more than two epitopes. Finally, we assume that at time \( T_{\text{sample}} \) we sample \( n \) infected cells for their infecting HIV sequences. We will take various values for \( T_{\text{sample}} \), but we set \( n = 20 \), a typical value found in experimental studies.

We categorize HIV infected CD4 cells into four types. CD4 cells infected by an HIV genome possessing both \( e_1 \) and \( e_2 \) will be founder type infected cells, or more succinctly founder cells. CD4 cells with HIV genomes possessing only \( e_1 \) will be called m2 cells (m2 stands for missing \( e_2 \)) and similarly m1 cells contain HIV genomes with only \( e_2 \). CD4 cells possessing neither epitope will be m12 cells.

Let \( F(t), M_1(t), M_2(t) \) and \( M_{12}(t) \) be the number of founder, m1, m2, and m12 cells at time \( t \), respectively. Since we start with one founder cell, we have \( M_1(0) = M_2(0) = M_{12}(0) = 0 \) and \( F(0) = 1 \). Up to \( T_{\text{peak}} \) all infected cells are assumed equally fit. More precisely, we think of infected cells as both dying and giving birth to newly infected cells at a certain rate. Corresponding to our choice of time scale, infected cells die at rate 1. For \( t < T_{\text{peak}} \) we assume that an infected cell gives birth to a newly infected cell at rate \( 1 + r(t) \), regardless of the infecting HIV genome. To be clear, for any time interval \([t, t + \Delta t]\) there is a probability approximately \( \Delta t \) that an infected cell will die and a probability \( (1 + r(t))\Delta t \) that the infected cell will produce a child infected cell.

We will always take \( r(t) > 0 \) reflecting the increase in infected CD4 cell
population size during early acute stage. We consider two specific instances of \( r(t) \): 

\[
\begin{align*}
  r(t) &= r_0 \quad \text{constant model,} \\
  r(t) &= 2r_0 - t\left(\frac{2r_0}{T_{\text{peak}}}\right) \quad \text{linear model,}
\end{align*}
\]

where \( r_0 = \frac{\log(N)}{T_{\text{peak}}} \). In the formula for \( r_0 \), \( N \) represents a constant that controls the number of infected cells at \( T_{\text{peak}} \). As we will show in section 2, the form of \( r_0 \) guarantees that at time \( T_{\text{peak}} \) the infected cell population will be order \( N \).

Estimates for the number of infected cells during chronic infection are in the range \( 10^6 - 10^8 \), see [10] and references therein. We take the upper range since we are considering peak infection and so we set \( N = 10^8 \). Intuitively, the constant model assumes an HIV population with no carrying capacity constraints while the linear model assumes the HIV population reaches a carrying capacity near \( N \) at \( T_{\text{peak}} \).

At \( T_{\text{peak}} \), CD8 cells with TCRs binding to either \( e_1 \) or \( e_2 \) begin to activate and proliferate. We assume that the resulting CTLs kill infected cells possessing \( e_1 \) with rate \( \lambda_1 \) and infected cells possessing \( e_2 \) with rate \( \lambda_2 \). Infected cells possessing both epitopes, i.e. founder cells, are killed at rate \( \lambda_1 + \lambda_2 \). We do not model the proliferation of CTLs as there is no general consensus for how such proliferation occurs [2]. Rather, we simply assume that at time \( T_{\text{peak}} \) the CTL attack is switched on and stays on up to time \( T_{\text{sample}} \).

We assume that after \( T_{\text{peak}} \), the total infected cell population size remains fixed. To do this, we assume every death event is associated with a birth event, i.e. a Moran model [8]. More precisely, infected cells die at rate 1 irrespective of CTL attack, and at rates \( \lambda_1 \), \( \lambda_2 \) due to CTL attack at the epitopes \( e_1 \) and \( e_2 \) respectively. When a death event occurs, we uniformly select an existing infected cell to produce an offspring infected cell that replaces the dying cell. Figure 1 summarizes the model.

We let \( \mu \) be the HIV mutations rate per base pair per cell infection and set \( \mu = 2 \times 10^{-5} \) [23]. We will distinguish between two types of mutation. A mutation that changes \( e_1 \) or \( e_2 \) so that the epitope is no longer present in the HIV genome will be referred to as an epitope mutation, all other mutations will be referred to as neutral mutations. We assume \( e_1 \) and \( e_2 \) can both be removed by single base pair mutations.

### 1.2 Sampling Genealogies

The model we have described is stochastic for two reasons. First, infected cells create offspring stochastically according to the birth and death rates specified above. Second, epitope mutations occur with probability \( \mu \) and so the number of, say, founder cell to m1 cell mutations in any time period is stochastic. To sample the genealogy of \( n \) HIV infected cells sampled at \( T_{\text{sample}} \), we first sample a given realization of \( F(t), M_1(t), M_2(t), M_{12}(t) \) for \( t \in [0, T_{\text{sample}}] \). We then sample a genealogy conditioned on these dynamics. Other authors have used
a similar approach. For instance, in [12, 9] the authors sampled genealogies from a strong selective sweep and in [4] the authors sampled genealogies from a population under selection-mutation equilibrium.

However, there are important differences between our sampler and that of previous authors. The selective sweep results of previous authors are classical sweeps in that a single advantageous allele arises and sweeps through the population. Due to high mutation rates and large population size, HIV selective sweeps occur due to many mutations that produce numerous copies of a selected allele, such sweeps have been termed soft sweeps [11]. Further, previous authors that have considered recurrent mutation have always assumed selection-mutation equilibrium. This is an important technical assumption as it makes the stochastic processes considered time reversible. In our case, the stochastic processes considered are not reversible because acute HIV infection is not in equilibrium. In practice this significantly raises the computational hurdles required for sampling.

The genealogies sampled depend on the CTL attack at e1 and e2. Consequently, the genealogies depend on epitope mutations and this is why we track the dynamics of the four cell types. However, neutral mutations do not affect the genealogies and so neutral mutations can be placed on top of the genealogy sampled according to whatever mutation model one wishes to consider, e.g.

Figure 1: Summary of Model
infinite sites model, JC69, HKY85, etc. Our genealogies provide a means to understand genetic diversity throughout the HIV genome conditioned on CTL attack at a given pair of epitopes.

We now proceed to describe the genealogy sampler we construct. As we do so, we will also derive some theoretical results describing the genealogy distribution. In section 2 we consider the time period \([0, T_{\text{peak}}]\). As a warm up to our full problem, and also due to its own interest, we consider a situation in which we sample at \(T_{\text{peak}}\) rather than \(T_{\text{sample}}\). In this setting we describe the corresponding genealogy sampler and the properties of the underlying genealogy distribution. Since CTL attack occurs at \(T_{\text{peak}}\), we refer to \([0, T_{\text{peak}}]\) as the pre-attack period and genealogies formed from samples at \(T_{\text{peak}}\) as pre-attack genealogies. In section 3 we consider our full problem of sampling at \(T_{\text{sample}}\). We refer to the time interval \([T_{\text{peak}}, T_{\text{sample}}]\) as the post-attack period and genealogies formed from samples at \(T_{\text{sample}}\) as post-attack genealogies. Finally, in section 4 we discuss two applications of our sampler and mention some future work.

## 2 Pre-attack Genealogies

In this section we consider the genealogies that arise when we sample \(n\) infected cells at \(T_{\text{peak}}\). As we are interested in constructing genealogies, all mutations considered will be epitope mutations. In sections 2.1-2.3 we consider the genealogies formed when one samples infected cell exclusively of one type, i.e. founder, m1, m2, or m12 cell types. Then in section 2.4 we combine the results of the previous sections to specify a sampler for arbitrary combinations of sample cell types.

Our approach, used repeatedly in each of the subsections below, is to split the dynamics into stochastic and deterministic time periods. Roughly, when the number of infected cells of a certain type is small, the population dynamics of the cell type are stochastic and must be modeled with a branching process. On the other hand, when the number of infected cells is large, the dynamics are deterministic and can be modeled with an ODE. Splitting the dynamics into these two periods is advantageous because solving the relevant ODEs is computationally less expensive than simulating the branching processes. This observation is not new and has been exploited by many authors, see for example [32, 28]. Importantly, the decomposition into stochastic and deterministic periods serves to simplify the sampling of lineages. As we shall show, during a deterministic period coalescent events rarely happen. Sampling genealogies is then essentially reduced to sampling lineages during stochastic periods.

Throughout the rest of this paper, we make constant use of results from branching process theory. We consider continuous time, branching processes. An introduction to the theory of such processes can be found in Chapter 3 of [3]. Mostly we depend on moment estimates for the number of individuals in a branching process at a given time, these results can be found in section 4 of chapter 3 in [3].
2.1 founder cell sampling

We first describe how we sample the number of founder cells, \( F(t) \). Initially \( F(t) \) behaves stochastically, indeed the variance of \( F(t) \) is of relative order \( O(F(t)) \) for small \( t \). However, due to averaging effects in the number of births and deaths, as \( F(t) \) becomes large its dynamics become roughly deterministic. To make this more precise, consider \( F(T_{\text{peak}}) \) conditioned on \( F(t) \), denoted \( F(T_{\text{peak}}) \bigg|_{F(t)} \), for some \( t < T_{\text{peak}} \). Under our model, each of the \( F(t) \) infected cells at time \( t \) evolves according to a branching process. Let \( X(T_{\text{peak}}, t) \) be the number of descendants at time \( T_{\text{peak}} \) produced by a single infected cell at time \( t \). A central limit theorem argument (hereafter CLT) gives,

\[
F(T_{\text{peak}}) \bigg|_{F(t)} \approx F(t)E[X(T_{\text{peak}}, t)] \left( 1 + \frac{\sqrt{V[X(T_{\text{peak}}, t)]}}{E[X(T_{\text{peak}}, t)]} \right) \sqrt{\frac{1}{F(t)}N(0, 1)},
\]

(2)

where \( N(0, 1) \) is a standard normal. Standard branching process computations, see [3], show \( \frac{\sqrt{V[X(T_{\text{peak}}, t)]}}{E[X(T_{\text{peak}}, t)]} = O(1) \) and so the relative error of approximating \( F(T_{\text{peak}}) \bigg|_{F(t)} \) by \( E[F(T_{\text{peak}}) \mid F(t)] \) is \( O(\frac{1}{\sqrt{F(t)}}) \) (see, for example, [32] for a similar analysis). Monte carlo computations with \( N \) set to \( 10^6 \) show that if we choose \( F(t) = 1000 \) then the average relative error in the approximation of \( F(T_{\text{peak}}) \bigg|_{F(t)} \) by \( E[F(T_{\text{peak}}) \mid F(t)] \) is \( 4\% \). Monte carlo computations for \( N = 10^8 \) took too long to run for an accurate estimation of error to be computed, but the error for \( N = 10^6 \) should be approximately that of \( N = 10^6 \). With these results in mind, if we set \( f(t) = \frac{F(t)}{N} \) then once \( F(t) > 1000 \) we have the following dynamics which have relative error of less than \( 4\% \),

\[
\frac{df}{dt}(t) = r(t)f(t).
\]

(3)

With (3) in mind, we split the sampling of \( F(t) \) into a stochastic period \([0, T_f]\) and a deterministic period \([T_f, T_{\text{peak}}]\). We define \( T_f \) as the first time \( t \) for which \( F(t) = 1000 \) where \( F(t) \) is evolved stochastically up to \( T_f \). This is essentially the approach used in [13]. On \([0, T_f]\) we track \( F(t) \) as a branching process, storing the time and nature of all jumps. At \( T_f \) we set \( f(T_f) = \frac{F(T_f)}{N} \) and then use (3) to solve for the dynamics of \( f(t) \). Standard results in branching processes show that \( E[f(T_{\text{peak}})] = 1 \) and that the variance is \( O(1) \), demonstrating that \( N \) is indeed the order of infected founder cells at \( T_{\text{peak}} \).

Consider sampling \( n \) founder cells at \( T_{\text{peak}} \). We first demonstrate that lineages formed from our samples rarely coalesce in the deterministic time period \([T_f, T_{\text{peak}}]\). Let \( T_{\text{coal}}^{(n)} \) be the time at which the \( n \) lineages formed from the samples coalesce into \( n - 1 \) lineages and consider \( t \) for which \( t > T_f \) and \( t > T_{\text{coal}}^{(n)} \). There are \( n \) separate lineages at \( t \) and each lineage can be associated with a
founder cell alive at time t, collectively we refer to these n founder cells as the lineage cells. Suppose that a birth event occurs at time t. Then there will be a coalescent event if some pair of the lineage cells are the parent and child corresponding to the birth event. By symmetry this has probability \( \frac{n(n-1)}{F(t)(F(t)-1)} \).

In the time interval \([t, t + \Delta t]\) the number of birth events will be approximately \( F(t)(1 + r(t))\Delta t \), and a CLT argument similar to the one above shows that this has high relative accuracy if \( F(t)\Delta t \) is large. Assuming this, the probability of a coalescent event in \([t, t + \Delta t]\) is approximately \( \frac{n(n-1)(1+r(t))}{F(t)}\Delta t \).

Since \( F(t) \) is guaranteed large when \( t > T_f \), we can integrate over many such intervals to find,

\[
P(T_{\text{coal}}^{(n)} < T_f) \approx \exp\left[ -n(n - 1) \int_{T_f}^{T_{\text{peak}}} ds \frac{1 + r(s)}{Nf(s)} \right].
\]

(4)

Notice that the right side of (4) is deterministic since \( f(T_f) = \frac{1000}{N} \). With \( n = 20 \), for both the constant and linear model, (4) gives \( P(T_{\text{coal}}^{(n)} < T_f) \approx 0.99 \), showing that on \([T_f, T_{\text{peak}}]\) the lineages stay separated with high probability. We emphasize that (4) only holds when \( t > T_f \) since this restriction insures that \( F(t) \) is large. Indeed this is precisely why we need to track the jumps of \( F(t) \) prior to \( T_f \).

To sample genealogies, we sample \( n \) founder cells at time \( T_f \) and use the dynamics of \( F(t) \) on \([0, T_f]\) to form the sampled genealogies between \([0, T_{\text{peak}}]\). Since we track every change in value of \( F(t) \) on \([0, T_f]\) we can explicitly consider each birth event and compute the corresponding probability that two lineages coalesce. Between time \( T_f \) and \( T_{\text{peak}} \) our genealogies are simply \( n \) separate lineages.

In [34], Slatkin and Hudson showed that sampling from a population that has undergone exponential growth results in star-like genealogies. Our discussion above reflects this phenomenon, although we are not assuming deterministic exponential growth.

2.2 m1 and m2 cell sampling

We now consider sampling of m1 cells or m2 cells at \( T_{\text{peak}} \). Since we assume that prior to \( T_{\text{peak}} \) all infected cells are equally fit, for concreteness we phrase our arguments solely in terms of m1. As we will show through the results of section 2.3, most m1 cells sampled at time \( T_{\text{peak}} \) possess only a single epitope mutation in their lineage. For such lineages there exists a time \( t \) for which on \([0, t]\) all cells in the lineage are founder cells, while on \([t, T_{\text{peak}}]\) all cells in the lineage are m1 cells. An m1 cell is born at time \( t \) that has a founder cell parent and we refer to such m1 cells as initial m1 cells. Each m1 cell at \( T_{\text{peak}} \) is associated with exactly one initial m1 cell, although different m1 cells may and often do share the same initial m1 cell. The occurrence of multiple initial m1 cells reflects the soft sweep dynamics of HIV selection that we alluded to in section 1.1.

The distribution of \( M_1(T_{\text{peak}}) \), the number of m1 cells at \( T_{\text{peak}} \), is described by Luria-Delbrück distributions which have been studied extensively, see [38] for an excellent review and [16, 26, 39] for more recent results. However, since
we need to build lineages on the dynamics of $M_1(t)$, knowing the distribution of $M_1(T_{\text{peak}})$ is insufficient. Indeed, in this subsection our novel result is a method for efficiently sampling genealogies from a population evolving according to a Luria-Delbrück process.

As we did for $F(t)$, we divide the dynamics of $M_1(t)$ prior to $T_{\text{peak}}$ into a stochastic period and a deterministic period. To precisely define these two time periods, set $T_I$ so that $f(T_I) = A N \mu$ where $A$ is a tuning parameter for our sampler. Increasing $A$ will produce a more accurate sampler but at greater computational expense. We let $[0, T_I]$ be the stochastic period and $[T_I, T_{\text{peak}}]$ the deterministic period.

To explain our approach in this subsection we introduce some additional notation. Let $D_{\text{det}}$ and $D_{\text{stoch}}$ be the number of $m_1$ cells at $T_{\text{peak}}$ that descend from initial $m_1$ cells born during the deterministic and stochastic periods respectively ($D_{\text{det}}$ and $D_{\text{stoch}}$ will be defined rigorously below). We have

$$M_1(T_{\text{peak}}) = D_{\text{det}} + D_{\text{stoch}}$$

since every $m_1$ cell must descend from an initial $m_1$ cell. Previous results involving Luria-Delbrück distributions give \cite{16, 26},

$$M_1(T_{\text{peak}}) \approx QS + Q \log(Q), \quad (5)$$

where

$$Q = \int_0^{T_{\text{peak}}} dt \mu F(t), \quad (6)$$

and $S$ is a stable distribution with $\alpha = 1$, skewness parameter $\beta = 1$ and scale parameter $c$ dependent on the form of $r(t) \ [27]$. For example, in the case of the constant model, i.e. $r(t) = r_0$, $c = \pi^2$. The relative error of the approximation (5) goes to 0 as $Q \to \infty$.

In subsection 2.2.1, we prove that $D_{\text{det}} \approx \mu N(T_{\text{peak}} - T_I)$ which gives,

$$D_{\text{stoch}} \approx QS + Q \log(Q) - \mu N(T_{\text{peak}} - T_I). \quad (7)$$

Further, we show that lineages rarely coalesce during the deterministic period. Sampling genealogies and $M_1(T_{\text{peak}})$ is then reduced to considering coalescent events during the stochastic period and sampling $D_{\text{stoch}}$. In subsection 2.2.2, we specify how to sample $D_{\text{stoch}}$ and in subsection 2.2.3 we consider coalescent events during the stochastic period.

### 2.2.1 Deterministic Period

Split the deterministic time period into intervals of size $\Delta t$. Let $D_k$ be the number of $m_1$ cells at time $T_{\text{peak}}$ that descend from initial $m_1$ cells born in time interval $Ivl_k = [s_k, s_{k+1}]$ where $s_k = T_I + k \Delta t$. We precisely define $D_{\text{det}}$ by $D_{\text{det}} = \sum_k D_k$. Let $I_k$ be the number of initial $m_1$ cells born in $Ivl_k$. $I_k$ is a Poisson r.v with mean,

$$E[I_k] \approx \mu F(s_k) \Delta t. \quad (8)$$

Let $X_k = X(T_{\text{peak}}, s_k)$, the number of descendants at time $T_{\text{peak}}$ produced by a single initial $m_1$ cell born in $Ivl_k$. Using standard formulas for the moments
of a branching process (see section 4 of chapter 3, equations (4) and (5) in [3]) we find

\[ E[X_k] = \frac{F(T_{\text{peak}})}{F(s_k)} = \exp\left(\int_{s_k}^{T_{\text{peak}}} ds \alpha(s) \exp\left(-\int_{s_k}^{s} ds' r(s')\right)\right), \]  

(9)

\[ E[X_k^2] = E[X_k]^2 \left(1 + \int_{s_k}^{T_{\text{peak}}} ds \alpha(s) \exp\left(-\int_{s_k}^{s} ds' r(s')\right)\right), \]

where \( \alpha(s) = 2 + r(s) \).

From the moments for \( X_k \) and \( I_k \), we find that the mean of \( D_k \) is constant over \( k \),

\[ E[D_k] = E[I_k]E[X_k] = \mu N \Delta t, \]  

(10)

Further, we find that \( D_{\text{det}} \) is well approximated by its mean,

\[ E[D_{\text{det}}] = \mu N (T_{\text{peak}} - T_f), \]  

(11)

\[ \frac{V[D_{\text{det}}]}{E[D_{\text{det}}]^2} \leq \frac{\sup_{\tau \leq T_{\text{peak}}} \alpha(t)}{A(T_{\text{peak}} - T_f)}. \]

For large \( A \) then, \( D_{\text{det}} \) is deterministic up to small error. Table 1 gives numerical results showing the accuracy of approximating \( D_{\text{det}} \) by \( E[D_{\text{det}}] \) for the cases \( A = 10, 50, 100 \) under both constant and linear models. All figures were produced by running 1000 simulations and averaging. The third column in the table gives values of \( E[D_{\text{det}}] \) while the fourth column gives the mean relative error of approximating \( D_{\text{det}} \) by \( E[D_{\text{det}}] \). As the table shows, raising \( A \) lowers \( E[D_{\text{det}}] \) since we are implicitly raising \( T_f \). Further, raising \( A \) improves the accuracy of the approximation \( D_{\text{det}} \approx E[D_{\text{det}}] \). Note that even for \( A = 10 \) the relative error is small.

To understand the effect of the deterministic time period on genealogies we consider two m1 cells sampled at \( T_{\text{peak}} \). By the results of section 2.1, we know that once an m1 sample lineage is composed of founder cells, it will rarely coalesce with any m1 or founder sample cell lineage prior to \( T_f \). To coalesce in the deterministic period then, two sample m1 cell lineages must descend from the same initial m1 cell. To compute this probability let \( \zeta_1, \zeta_2 \) be the birth times

<table>
<thead>
<tr>
<th>model</th>
<th>( A )</th>
<th>( E[D_{\text{det}}] )</th>
<th>( E\left[\frac{E[D_{\text{det}}] - D_{\text{det}}}{D_{\text{det}}} \right] )</th>
<th>formula (16)</th>
<th>error of (16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>10</td>
<td>11500</td>
<td>.03</td>
<td>.01</td>
<td>.2</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>8000</td>
<td>.025</td>
<td>.005</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>6500</td>
<td>.021</td>
<td>.003</td>
<td>.01</td>
</tr>
<tr>
<td>linear</td>
<td>10</td>
<td>21400</td>
<td>.013</td>
<td>.005</td>
<td>.26</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>17900</td>
<td>.006</td>
<td>.002</td>
<td>.014</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>16100</td>
<td>.005</td>
<td>.0015</td>
<td>.006</td>
</tr>
</tbody>
</table>

Table 1: Error for Deterministic Period Approximations
of the initial $m_1$ cells associated with the two sampled cells. We then consider $\zeta_1, \zeta_2$ conditioned on both samples being descendants of initial $m_1$ cells born in the deterministic period, labeling these conditioned versions as $\zeta_{1,\text{det}}, \zeta_{2,\text{det}}$.

There are two quantities we will require for construction of genealogies. The probability the samples share an initial $m_1$ cell, $P(\zeta_{1,\text{det}} = \zeta_{2,\text{det}})$, and the conditional density of initial $m_1$ cell birth time.

We first compute $P(\zeta_{1,\text{det}} = \zeta_{2,\text{det}} | \zeta_{1,\text{det}}, \zeta_{2,\text{det}} \in Ivl_k)$, where $X_k$ is the conditional density of initial $m_1$ cell birth time.

$$P(\zeta_{1,\text{det}} = \zeta_{2,\text{det}} | \zeta_{1,\text{det}}, \zeta_{2,\text{det}} \in Ivl_k) = \sum_{i=1}^{I_k} \frac{X_{k,i}(X_{k,i} - 1)}{D_k(D_k - 1)}$$ (13)

If we now substitute $D_k = \sum_{i=1}^{I_k} X_{k,i}$ we can apply law of large number (hereafter LLN) and CLT arguments by dividing the numerator and denominator above by \( \frac{1}{T_k^2} \) to give,

$$P(\zeta_{1,\text{det}} = \zeta_{2,\text{det}} | \zeta_{1,\text{det}}, \zeta_{2,\text{det}} \in Ivl_k) \approx \frac{1}{T_k^2} \left( E[X_k^2] - E[X_k] \right) \left( 1 + O\left( \frac{1}{\sqrt{X_k}} \right) \right) \left( 1 + O\left( \frac{1}{\sqrt{X_k}} \right) \right)$$ (14)

Now by symmetry we have $P(\zeta_{1,\text{det}} \in Ivl_k) = \frac{D_k}{D_{\text{det}}} \approx \frac{E[X_k]E[I_k]}{E[D_{\text{det}}]}$. Combining this approximation with (8), (9), and (14) gives

$$P(\zeta_{1,\text{II}} = \zeta_{2,\text{II}}) \approx \sum_k \frac{\Delta t}{\mu N(T_{\text{peak}} - T_1)^2} \left( \frac{E[X_k^2]}{E[X_k]} - 1 \right).$$ (15)

Rigorously, if we take $\Delta t \to 0$ and $A \to \infty$ such that $A \Delta t$ remains large, we have

$$P(\zeta_{1,\text{II}} = \zeta_{2,\text{II}}) \approx \int_{T_1}^{T_{\text{peak}}} ds \frac{1}{\mu N(T_{\text{peak}} - T_1)^2} \left( \frac{E[X_s^2]}{E[X_s]} - 1 \right)$$ (16)

where $X_s = X(T_{\text{peak}}, s)$. For finite $A$, (16) becomes an approximation with an error of $O(\frac{1}{\sqrt{A}})$. Generally, (16) can be evaluated numerically. For the constant model, if we set $T_1$ by $E[M_1(T_1)] = \frac{A}{\mu}$ rather than $M_1(T_1) = \frac{A}{\mu}$, we have the analytic approximation,

$$P(\zeta_{1,\text{II}} = \zeta_{2,\text{II}}) \approx \frac{2(1 + r_0)}{A \log(\frac{X_s}{\mu})}. \tag{17}$$
Notice that the time of mutation is distributed. Let the time of the initial \(m_1\) cell birth is uniformly distributed on \([0, 10]\), but collapses for \(A = 50, 100\). The small values given in Table 1 for (16) consider only two samples. When we consider \(n = 20\) the values in Table 1 for (16) apply to all 95 possible pairs of samples, giving a more significant probability that a coalescent events will occur. Still, we see that the deterministic period will produce few coalescent events unless \(n\) is large.

Turning finally to \(dP(\zeta_{1, \text{det}} = t \mid \zeta_{1, \text{det}} = \zeta_{2, \text{det}})\), a similar analysis gives

\[
dP(\zeta_{1, \text{det}} = \zeta_{2, \text{det}} = t \mid \zeta_{1, \text{det}} = \zeta_{2, \text{det}}) \approx \frac{E[X_t^2]}{E[X_t]} - 1 \int_{T_1}^{T_{\text{peak}}} ds \left( \frac{E[X_t^2]}{E[X_t]} - 1 \right).
\]

(18)

For the constant model we find,

\[
dP(\zeta_{1, \text{det}} = \zeta_{2, \text{det}} = t \mid \zeta_{1, \text{det}}) \approx r_0 \exp[-r_0(t - T_1)].
\]

(19)

### 2.2.2 Stochastic Time Period

In the stochastic time period, \([0, T_1]\), there will be few initial \(m_1\) cells. Let \(I_{\text{stoch}}\) be the number of initial \(m_1\) cells in the stochastic time period. \(I_{\text{stoch}}\) is Poisson distributed with mean \(\int_0^1 dt \mu F(t)\), under the constant model \(E[I_{\text{stoch}}] \approx \frac{A}{\tau}\). We let \(X_{\text{stoch}}\) be the number of \(m_1\) cells at \(T_{\text{peak}}\) descendant from a single initial \(m_1\) cell in the stochastic period and \(X_{\text{stoch}, i}\) for \(i = 1, 2, \ldots, I_{\text{stoch}}\) iid copies. The time of the initial \(m_1\) cell birth is uniformly distributed on \([h([0, T_{\text{stoch}}]), U h(T_1)])\) where \(h(t) = \int_0^t ds \mu F(s)\) and if we map back using \(h^{-1}\) we can determine how the time of mutation is distributed. Let \(t^*\) be the time of the mutation and \(U\) a uniform r.v. on \([0, 1]\).

\[
t^* = h^{-1}(U h(T_1)).
\]

(20)

To sample \(X_{\text{stoch}}\) exactly, we sample \(t^*\) through (20) and then sample the branching process \(X(T_{\text{peak}}, t^*)\). This is not computationally burdensome as in the constant model we sample \(X_{\text{stoch}}\) approximately \(\frac{A}{\tau}\) times and in the linear model we expect to require fewer samples.

We can precisely define \(D_{\text{stoch}}\) by \(D_{\text{stoch}} = \sum_{i=1}^{I_{\text{stoch}}} X_{\text{stoch}, i}\), and so by sampling \(X_{\text{stoch}}\) we are able to sample \(D_{\text{stoch}}\). Expressing \(M_1(T_{\text{peak}})\) as \(M_1(T_{\text{peak}}) = D_{\text{det}}(1 + D_{\text{stoch}}/D_{\text{det}})\), we can examine the stochasticity of \(M_1(T_{\text{peak}})\) by considering the ratio \(D_{\text{stoch}}/D_{\text{det}}\). We have shown \(D_{\text{det}} \approx \mu N(T_{\text{peak}} - T_1)\). Recalling (6), we have

\[
\frac{Q}{D_{\text{det}}} \approx \frac{1}{T_{\text{peak}} - T_1} \int_0^{T_{\text{peak}}} dt f(t) \approx \tilde{Q}.
\]

(21)

where

\[
\tilde{Q} = \frac{1}{T_{\text{peak}} - T_1} \int_{T_1}^{T_{\text{peak}}} dt f(t).
\]

(22)

The second approximation in (21) follows from \(f(t) \leq \frac{A}{\mu N} \ll 1\) for \(t < T_1\). Notice that \(Q \leq 1\) since \(f(t) \leq 1\). Under the constant model, \(\tilde{Q} \approx \frac{1}{\log(\mu N/A)}\).
From (7) and (21), we find

$$\frac{D_{\text{stoch}}}{D_{\text{det}}} \approx \tilde{Q}(S + \log(Q)) - 1$$  \hspace{1cm} (23)

(23) seems to suggest the possibility that $D_{\text{stoch}}$ is negative, however for large $Q$ the approximations we have used become more precise and the ratio $D_{\text{stoch}}/D_{\text{det}}$ will be positive. In the setting of HIV, $Q$ is large. For example, under the constant model with $r_0 = .5$, $Q = 4000$.

Figure 2 shows the probability distribution of $\frac{D_{\text{stoch}}}{D_{\text{det}}}$ compared with the approximation (23) for the constant model with $A = 100$. Results were derived by running 1000 simulations and producing the histogram shown in the figure. As can be seen, (23) is a fairly good approximation. Further $\frac{D_{\text{stoch}}}{D_{\text{det}}}$ has heavy tails and so with significant probability $D_{\text{stoch}}$ dominates $D_{\text{det}}$. When $A = 100$, $\frac{D_{\text{stoch}}}{D_{\text{det}}}$ is somewhat skewed as $T_1$ is large. To see that the heavy tails of $\frac{D_{\text{stoch}}}{D_{\text{det}}}$ are not an artifice of our choice $A = 100$, Figure 3 shows $\frac{D_{\text{stoch}}}{D_{\text{det}}}$ for $A = 1$ and here too we see that the heavy tails allow $D_{\text{stoch}}$ to often dominate $D_{\text{det}}$.

### 2.2.3 sampling m1 and m2 genealogies

Consider sampling $n$ m1 cells at $T_{\text{peak}}$. To sample a genealogy we first determine whether each sampled cell descends from an initial m1 cell in the deterministic or the stochastic time period. To do this we sample $D_{\text{stoch}}$ by sampling $I_{\text{stoch}}$ and setting $D_{\text{stoch}} = \sum_{i=1}^{I_{\text{stoch}}} X_{\text{stoch},i}$. Then a sampled m1 cell has an associated initial m1 cell that is in the deterministic or stochastic period with probabilities $1 - \frac{D_{\text{stoch}}}{D_{\text{det}}}$ and $\frac{D_{\text{stoch}}}{D_{\text{det}}}$ respectively.

Having separated the m1 sample cells into a deterministic group, i.e. those sample cells that descend from initial m1 cells born in the deterministic period, and a stochastic group, we now consider coalescent events within each group. The probability of any pair of samples from the deterministic period group coalescing and the time of coalescence was computed in section 2.2.1. The probability of three sampled m1 cells coalescing in the deterministic period is lower order and can be ignored if $nP(\zeta_{1,\text{det}} = \zeta_{2,\text{det}})$ is small, a situation that holds for our parameter choices. To determine the number of coalescent events in the deterministic time period we draw a binomial random variable with success probability given by $P(\zeta_{1,\text{det}} = \zeta_{2,\text{det}})$ and $\binom{n}{2}$ trials. Since $nP(\zeta_{1,\text{det}} = \zeta_{2,\text{det}})$ is small, the probability of choosing more than $\frac{n}{2}$ coalescent events is very small. For example, in the constant model with $A = 10$ and $n = 20$ we find that the probability of choosing more than 10 coalescent events is $4 \times 10^{-6}$. Given the number of coalescent events in the deterministic period, we uniformly choose pairs of samples from the deterministic group to coalesce.

For the sample cells in the stochastic group, a sample descends from the initial m1 cells associated with $X_{\text{stoch},j}$ with probability $\frac{X_{\text{stoch},j}}{D_{\text{stoch}}}$. We can then assign all stochastic period samples to initial m1 cells. The genealogy of all samples associated with $X_{\text{stoch},j}$, say, can then be sampled up to the start time of $X_{\text{stoch},j}$ since we track the jumps of the branching process explicitly.
Once all m1 sampled cells are traced back to their initial m1 cells, the lineages are all composed of founder cells and we may apply the methods of section 2.1.

2.3 m12 sampling

m12 cells can arise either through two epitope mutations both occurring in a single founder cell, a situation we will refer to as a double epitope mutation, or by one epitope mutation occurring in an m1 or m2 cell. In this section we show that the number of m12 cells at $T_{\text{peak}}$ will be small and often will be zero.

Double epitope mutations are produced at rate $\mu^2 N f(t)$. The number of double mutations occurring by time $T_{\text{peak}}$, which we denote as $I_{\text{m12, double}}$, has expected value $\mu^2 \int_0^{T_{\text{peak}}} ds f(s)$ which evaluates to .03 and .11 in the constant and linear model respectively. Further, the expected number of double epitope mutations prior to time $t$ collapses exponentially as $t$ gets smaller. However, as in the case of the stochastic time period, the distribution of the number of m12 cells at $T_{\text{peak}}$ that descend from double epitope mutations has a heavy tail.

A similar picture holds for m12 cells that arise through an epitope mutation in m1 or m2 cells. Figure 4 shows a histogram of $M_{12}(T_{\text{peak}})$ based on 10000 simulations. The first bar represents the case $M_{12}(T_{\text{peak}}) = 0$ and extends off the graph to a frequency of .58. As can be seen, typically $M_{12}(T_{\text{peak}})$ is quite modest but we see the long tails reflecting rare events in which an early m12 initial cell is created.

Since there are so few initial m12 cells, we track the occurrence of each initial m12 cell birth event and then the branching process that gives the number of its descendants up to $T_{\text{peak}}$. To produce genealogies from m12 samples we simply work backwards in time using the branching process until each sampled m12 cell’s lineage reaches an initial m12 cell. The parent of an initial m12 cell could be a founder, m1, or m2 cell. Regardless, sections 2.1 and 2.2 then show how we can extend the lineages down to $t = 0$.

2.4 Sampling Pre-Attack Genealogies

Consider now the general problem of sampling a genealogy based on an arbitrary choice for the cell types sampled at $T_{\text{peak}}$. In the previous sections we have described how to sample genealogies for sample cells of a single type. We simply combine these methods to sample a genealogy for sample cells of mixed type. If we sample according to the population sizes of the cell types at time $T_{\text{peak}}$, i.e. $F(T_{\text{peak}})$, $M_1(T_{\text{peak}})$, $M_2(T_{\text{peak}})$ and $M_{12}(T_{\text{peak}})$, all cells will be of founder type with high probability since $F(t) = O(N) \gg M_1(t) = O(\mu N)$. The result will be the star genealogies described in subsection 2.1.

Up to now we have not specified a choice for $A$. In the pre-attack case, genealogies can be sampled relatively quickly and so $A$ can be set quite high. In practice, for the monte carlo experiments considered in this section, we have found $A = 100$ to be a good choice. On our CPU, an Intel Pentium running at 3.2 GHz, a single tree is sampled in .078, .094 and .125 seconds with $A = 10, 50, 100$ respectively.
Figures 5-7 show specific sampled genealogies (all genealogy figures were generated using FigTree v1.3.1). Figure 5 represents cells sampled at $T_{\text{peak}}$ according to the cell population frequencies. As a result, all sampled cells are of founder type. The $f$ in the tip labels shows the cell to be of founder type while the number after the underscore simply labels the samples. The number above each branch gives the branch length in units of 2 days. Notice that the genealogy is star shaped. Figure 6 shows a sampled genealogy for which all sampled cells are m1 type (the m1 in the tip labels reflects this). In this case, the genealogy contains coalescent events throughout the time period $[0, T_{\text{peak}}]$ due to shared initial m1 cells. The genealogy in Figure 7 is sampled under the same assumptions as that of Figure 6. However, the numerous coalescent events in the upper portion of the genealogy reflect the birth of an initial m1 cell relatively early in the stochastic period. Due to its early birth time, this initial m1 cell produces many descendants that dominate the population of m1 cells at $T_{\text{peak}}$. The occurrence of such early initial m1 cell births corresponds to sampling $D_{\text{stoch}}$ from the tail of $D_{\text{det}}$, see Figures 2 and 3.

3 Post-Attack Genealogies

In this section, we describe how to sample from our full model. In section 3.1, we describe how to sample genealogies in the case of attack at a single epitope, say e1. This is actually just a special case of our full model in which we take $\lambda_2 = 0$, however highlighting this special case provides valuable insight into the relationship between CTL attack and HIV genetic diversity. In section 3.2 we describe the sampling of genealogies under our full model of double attack. Before continuing, we make a slight adjustment in our notation that will simplify our formulas. Let $\hat{N}$ be the number of infected cells at time $T_{\text{peak}}$. Then, for $t > T_{\text{peak}}$ we define $f(t)$ by $f(t) = \frac{F(t)}{\hat{N}}$. Thus, for $t > T_{\text{peak}}$, $f(t)$ represents the fraction of infected cells of founder type and this correspondence motivates our change of notation. Note that $E[\hat{N}] = N$. For simplicity, we will always take $\lambda_1 \geq \lambda_2$. Recall, that in the post-attack period the number of infected cells stays constant and evolves according to a Moran model.

3.1 Single Attack

In this section we restrict ourselves to CTL attack on e1 and so only $F(t)$ and $M_1(t)$ apply. We show that in the post-attack period, $[T_{\text{peak}}, T_{\text{sample}}]$, lineages rarely coalesce, thus reducing sampling at $T_{\text{sample}}$ to sampling at $T_{\text{peak}}$. To start on this path, first notice that at the outset of the post-attack period $F(T_{\text{peak}}) = O(N) \gg 1$ and $M_1(T_{\text{peak}}) = O(\mu N) \gg 1$. Consequently, post-attack dynamics are deterministic by essentially the same arguments used to justify (3). With this in mind, we can describe m1 cell dynamics through the ODE,

$$\frac{dm_1}{dt}(t) = \lambda_1 m_1(t)(1 - m_1(t)) + \mu f(t). \quad (24)$$
where \( m_1(t) = \frac{M_1(t)}{N} \). Plugging the equality \( f(t) + m_1(t) = 1 \) into (24) gives,

\[
\frac{dm_1}{dt}(t) = \lambda_1 m_1(t)(1 - m_1(t)) + \mu(1 - m_1(t)).
\] (25)

In the pre-attack period we showed that the coalescent probabilities during the deterministic period are small. In the post-attack case the coalescent probabilities are even smaller because the deterministic period starts with \( \mathcal{O}(\mu N) \) m1 cells already in existence. Indeed, essentially the same arguments used in section 2.1 to derive (4) may be used to show that the probability of no coalescent event in the post-attack period is given by,

\[
P(\text{no coal. event}) = \exp[-n(n - 1) \int_{T_{\text{sample}}}^{T_{\text{peak}}} ds \frac{1 + \lambda_1 f(s)}{Nm_1(s)}] \] (26)

Intuitively, since \( M_1(T_{\text{peak}}) > D_{\text{det}} = \mu N(T_{\text{peak}} - T_1) \), we have \( Nm_1(s) > \mu N(T_{\text{peak}} - T_1) \) and the probability of a coal event will be small. More precisely, (26) can be integrated explicitly. We find that under the constant model for \( \lambda_1 = .05, .1, .3, .5, 1 \) the probability of no coal event is .86, .88, .92, .93, .95 respectively. For the \( \lambda \) values mentioned, the probability of more than one coalescent event evaluates to .01, .008, .004, .003, .002 respectively. The probabilities of no coalescent events under the linear model are higher since more m1 cells exist at \( T_{\text{peak}} \), see Table 1. Up to the cost of ignoring a single coalescent event, we can then assume that lineages do not coalesce during the post-attack period. Sampling genealogies then follows from the methods of section 2.

### 3.2 Double Attack

To sample genealogies in the double attack case we follow the same line of reasoning used in the single attack case. First, by identical arguments used to justify (25), we can describe the dynamics of founder, m1, and m2 cells deterministically.

\[
\dot{m}_1(t) = m_1(t)(\lambda_1 - w(t)) + \mu f(t),
\]

\[
\dot{m}_2(t) = m_2(t)(\lambda_2 - w(t)) + \mu f(t),
\]

\[
\dot{f}(t) = -w(t)f(t),
\] (27)

where \( m_2(t) = \frac{M_2(t)}{N} \) and \( w(t) = \lambda_1 m_1(t) + \lambda_2 m_2(t) \). \( w(t) \) is the relative fitness of the population measured against a founder cell. Notice that if \( \lambda_2 = 0 \) then the (27) reduces to (25) as it should. Also, \( w(T_{\text{peak}}) \approx 0 \) showing that initially in the post-attack phase the dynamics of m1 and m2 cells are essentially decoupled. In the third equation of (27), the expression \(-2\mu f(t)\) representing founder cells that mutate into m1 or m2 cells should be present. Similarly, in the first and second equations there should be expressions corresponding to mutations that convert m1 or m2 cells back to founder cells. Since \( \mu \) is small, these terms can be ignored without significant effect.
At the outset of the post-attack period, we have \( f(T_{\text{peak}}) \approx 1, m_1(T_{\text{peak}}), m_2(T_{\text{peak}}) \approx O(\mu) \) and correspondingly \( w(t) = O(\mu) \). Initially then, \( f(t) \approx 1 \) will remain true while \( m_1 \) and \( m_2 \) cells populations will grow exponentially with rates \( \lambda_1 - O(\mu) \) and \( \lambda_2 - O(\mu) \), respectively. Eventually \( m_1(t) \) and \( m_2(t) \) will rise to \( O(1) \) levels thereby raising \( w(t) \) to \( O(1) \) levels as well. As a result, \( f(t) \) will drop at an exponential rate. Eventually \( f(t) \) will be driven to negligible levels. In the case \( \lambda_1 > \lambda_2 \), \( m_1 \) cells will eventually drive \( m_2 \) cells to negligible levels as well.

Now, consider the dynamics of \( M_{12}(t) \) and define \( \tau \) by \( M_{12}(\tau) = \frac{1}{\mu} \). Since \( m_2 \) cells are at a selective advantage we are guaranteed the existence and uniqueness of \( \tau \). Define \( m_{12}(t) = \frac{M_{12}(t)}{N} \). Then, restricting to \( t \in [\tau, T_{\text{peak}}] \), the same arguments that gave (24) result in

\[
dm_{12}(t) = m_{12}(t)(\lambda_1 + \lambda_2 - \bar{\omega}(t))dt + \mu(m_1(t) + m_2(t)) + \mu^2 f(t),
\]

where \( \bar{\omega}(t) = w(t) + (\lambda_1 + \lambda_2)m_{12}(t) \). (28) can be added to (27) to form the system,

\[
\begin{align*}
\dot{m}_{12} &= m_{12}(t)(\lambda_1 + \lambda_2 - \bar{\omega}(t)) + \mu(m_1(t) + m_2(t)) + \mu^2 f(t), \\
\dot{m}_1(t) &= m_1(t)(\lambda_2 - \bar{\omega}(t)) + \mu f(t), \\
\dot{m}_2(t) &= m_2(t)(\lambda_2 - \bar{\omega}(t)) + \mu f(t), \\
\dot{f}(t) &= -\bar{\omega}(t) \cdot f(t),
\end{align*}
\]

Now we consider the problem of sampling a genealogy assuming all sampled cells are of \( m_2 \) type. The general problem of mixed sample cell type will be considered in section 3.3. As we will show directly below, on \([\tau, T_{\text{sample}}]\) the lineages of \( m_2 \) cells rarely coalesce. On the other hand, during \([T_{\text{peak}}, \tau]\) the lineages do coalesce and we will need to perform an analysis similar to the stochastic-deterministic decomposition in section 2.2. Let \( P(\tau, T_{\text{sample}}) \) be the probability that no lineages coalesce in \([\tau, T_{\text{sample}}]\). Using the same arguments that gave (26), we have

\[
P(\tau, T_{\text{sample}}) = \exp[-n(n - 1) \int_\tau^{T_{\text{sample}}} ds \cdot \frac{1 + \lambda_1 + \lambda_2 - \bar{\omega}(s)}{Nm_{12}(s)}]
\]

Plugging in the bound \( Nm_{12}(s) > \frac{1}{\mu} \) and \( \tau \geq T_{\text{peak}} \) gives \( P(\tau, T_{\text{sample}}) \leq 0.08 \) for \( T_{\text{sample}} \leq 60 \). This bound holds for any \( \lambda_1, \lambda_2 \) and is not sharp for cases of substantial attack strength. For example, in the case \( \lambda_1 = .3, \lambda_2 = .1 \) we find \( P(\tau, T_{\text{sample}}) = .004 \) when \( T_{\text{sample}} = 60 \). Hence, with small error we ignore coalescent events on \([\tau, T_{\text{sample}}]\).

For the time period \([T_{\text{peak}}, \tau]\) we may sample \( m_2 \) lineages using the same approach we used in the pre-attack case. To explain this connection, consider the creation of \( m_2 \) cells by mutations in \( m_1 \) cells during the post-attack period. This is analogous to the pre-attack case in which we consider \( m_1 \) cells created by mutations in founder cells. However, the analogy is not exact as in the pre-attack case \( F(0) = 1 \) and all cells have an average growth rate \( r(t) \). In
contrast, during the post-attack case we consider \( M_1(T) \) which typically is much greater than one. Further m1 and m12 cells have average growth rates \( \lambda_1 - \tilde{w}(t) \) and \( \lambda_1 + \lambda_2 - \tilde{w}(t) \) respectively. Also, in the pre-attack case we evolve the system to \( T \) while in the post-attack case we are considering evolution to \( \tau \) which is a random variable depending on the system dynamics. However, we can handle these deviations using the same type of arguments introduced in section 2. Indeed, in analogy to the pre-attack case here we define \( T_1 \) by \( M_1(T_1) = \frac{A}{\mu} \). If \( M_1(T_{\text{peak}}) > \frac{A}{\mu} \) we set \( T_1 = T_{\text{peak}} \) and if \( M_1(\tau) < \frac{A}{\mu} \) we set \( T_1 = \tau \). The value of \( T_1 \) is unique because \( M_1(t) \) is increasing on \([T_{\text{peak}}, \tau]\). Having defined \( T_1 \), we then define \( s_k, D_k, I_k, D_{\text{det}}, \zeta_{1,\text{det}}, \zeta_{2,\text{det}} \) in the same way as defined for the pre-attack setting. In the pre-attack case we could derive analytic formulas for \( D_{\text{det}} \), \( P(\zeta_{1,\text{det}} = \zeta_{2,\text{det}}) \) and \( dP(\zeta_{1,\text{det}} = t \mid \zeta_{1,\text{det}} = \zeta_{2,\text{det}}) \), but in the post-attack case analytic formulas are not possible in general because (27) cannot be integrated analytically. However, we are able to numerically compute these expressions. For example, we have

\[
\begin{align*}
I_k & \approx \mu \tilde{N} m_1(s_k) \Delta t \\
D_k & \approx I_k E[X_k] \Delta t,
\end{align*}
\]

where \( X_k = X_{12}(t, s_k) \) and \( X_{12}(t, \tau) \) gives the number of descendants at \( \tau \) produced by a single m12 cell born at \( t \). \( X_{12}(\tau, t) \) is described by a branching process with death rate 1 and birth rate \( 1 + \lambda_1 + \lambda_2 - \tilde{w}(t) \). We can define \( D_{\text{det}} = \sum_k D_k \) and in analogy to (16) derive,

\[
P(\zeta_{1,\text{det}} = \zeta_{2,\text{det}}) = \sum_{k=1}^{T_1 - T_{\text{peak}}} \left( \frac{D_k}{D_{\text{det}}} \right)^2 \frac{1}{I_k} \left( \frac{E[X_k^2] - E[X_k]}{E[X_k]^2} \right)
\]

The accuracy of (32) is examined through numerical simulations in Table 2. The results are produced by averaging over 1000 simulations. We consider (32) assuming a constant model in the pre-attack period. In the post-attack period we study two types of attack: a strong attack for which \( \lambda_1 = .7, \lambda_2 = .5 \) and a weak CTL attack for which \( \lambda_1 = .1, \lambda_2 = .05 \). The third and fourth columns of Table 2 give (32) and its relative error from the true probability of coalescence, respectively. Notice that the probability of coalescence is higher under strong attack than weak attack. This is because under strong attack initial m12 cells born early in the deterministic period tend to produce more progeny than later initial m12 cells. This is not true in the pre-attack case as seen by (10). The ratio \( \frac{D_{\text{stoch}}}{D_{\text{det}}} \) is also affected by the strength of attack. In the pre-attack case, the simulations we ran to form Figure 3 for which \( A = 1 \) estimated \( P(\frac{D_{\text{stoch}}}{D_{\text{det}} < 1}) = .9 \). In the post-attack case we estimate \( P(\frac{D_{\text{stoch}}}{D_{\text{det}} < 1}) = .86 \) and \( P(\frac{D_{\text{stoch}}}{D_{\text{det}} < 1}) = .79 \) for the weak and strong attack case respectively, again taking \( A = 1 \). Under strong attack, initial m12 cells born during the stochastic period produce many more descendants than initial m12 cells born during the deterministic period due to their selective advantage over m1 cells. As a result, \( \frac{D_{\text{stoch}}}{D_{\text{det}}} \) has heavier tails under strong attack than weak attack.
A formula (32) error of formula (16)

<table>
<thead>
<tr>
<th>model</th>
<th>A</th>
<th>formula (32)</th>
<th>error of formula (16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>strong attack</td>
<td>10</td>
<td>.06</td>
<td>.33</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>.01</td>
<td>.07</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>.008</td>
<td>.04</td>
</tr>
<tr>
<td>weak attack</td>
<td>10</td>
<td>.03</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>.008</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>.005</td>
<td>.03</td>
</tr>
</tbody>
</table>

Table 2: Error for Post-Attack Deterministic Period Approximations

Sampling m12 lineages formed through epitope mutations in m2 cells proceeds essentially as in the case of epitope mutations in m1 cells. However, while we are guaranteed an increasing value for \( m_1(t) \) on \([T_{\text{peak}}, \tau]\), \( m_2(t) \) will initially rise and then may fall due to competition from fitter m1 cells. In such a setting we split \([T_{\text{peak}}, \tau]\) into three periods using \( T_I \) and \( T_{II} \) where \( T_I, T_{II} \) are the two solutions of \( M_2(t) = \frac{A}{\mu} \). Then \([T_{\text{peak}}, T_I]\) and \([T_{II}, \tau]\) are two stochastic time periods and \([T_I, T_{II}]\) forms the deterministic time period. Our methods extend to this variation on our standard case.

### 3.3 Sampling Post-Attack Genealogies

To sample a genealogy under double attack we evolve the population dynamics according to the ODE (29) up to \( T_{\text{sample}} \). We choose our sample types according to the frequencies specified by \( f(T_{\text{sample}}) \), \( m_1(T_{\text{sample}}) \), \( m_2(T_{\text{sample}}) \) and \( m_{12}(T_{\text{sample}}) \). In section 3.2 we considered sampling genealogies of m12 cells but not founder, m1, or m2 cells. We now show that lineages of founder, m1, and m2 cells rarely coalesce in \([T_{\text{peak}}, T_{\text{sample}}]\). Intuitively, if we sample such cells, then their populations must be large through \([T_{\text{peak}}, T_{\text{sample}}]\) and this makes coalescent events unlikely. To demonstrate this, consider two founder cell samples. Such samples will exist with probability greater than .01 only if

\[
(1 - f(T_{\text{sample}}))^n \leq .99 \tag{33}
\]

For \( n = 20 \) the above inequality requires \( f(T_{\text{sample}}) \geq 5 \times 10^{-4} = \frac{25\mu N}{N} \). Since \( f(t) \) is strictly decreasing on \([T_{\text{peak}}, T_{\text{sample}}]\) we have \( F(t) \geq 25\mu N = O(10^5) \) and by the results of section 2.1 we know that sample founder cell lineages will stay separated on \([T_{\text{peak}}, T_{\text{sample}}]\). A similar observation holds for m1 and m2 samples. Both \( m_1(t) \) and \( m_2(t) \) monotonically increase up to some time and then monotonically decrease. The arguments of section 3.1 show that in the monotonically increasing portion lineages will rarely coalesce and arguments identical to those just made for founder cell lineages show that coalescent events rarely occur in the monotonically decreasing portion.

To sample post-attack genealogies we build m12 lineages on \([T_{\text{peak}}, T_{\text{sample}}]\) according to the methods described in section 3.2. Lineages corresponding to founder, m1, or m2 cells remain uncoalesced on \([T_{\text{peak}}, T_{\text{sample}}]\). Once we trace the lineages down to \( T_{\text{peak}} \), we apply the methods of section 2.
Figures 8-11 show sample genealogies under a model of double attack. All four genealogies were sampled with $A = 50$. In the double attack setting, when sampling up to $T_{\text{sample}} = 60$, our CPU generated a single tree in .42, .69 and .97 seconds with $A = 10, 50, 100$ respectively.

Figures 8-10 represent the same realization of infected cell dynamics sampled at $T_{\text{sample}} = 60, 40, 30$ respectively. That is, we chose the constant model for pre-attack period and $\lambda_1 = \lambda_2 = .5$ for CTL attack rates and then sampled the infected cell dynamics once for $[0, 60]$. On top of this single realization of infected cell dynamics, we then sampled three genealogies assuming different times for $T_{\text{sample}}$. Figure 8 corresponds to $T_{\text{sample}} = 60$, a time by which the m12 cells have swept through the population. The tip labels give the cell type, sample number, and cell type of the first ancestor along the sample lineage that is not of the sample cell type. For example, the tip labeled m12_11_a1 corresponds to a sampled m12 cell, this sample was the 11th sampled cell and it arose from a mutation to an m1 cell ancestor. The numerous coalescent events that occur on the bottom portion of the genealogy correspond to an m1 cell ancestor that mutated into an m12 cell early in the post-attack period and whose descendants eventually dominated the population. Figure 9 corresponds to $T_{\text{sample}} = 40$, at this time the population percentages for founder, m1, m2, and m12 cells were .07, .45, .22, .23 respectively. The result is less coalescent events because m1 and m2 cells can only coalesce in the pre-attack period. Finally, Figure 10 represents sampling when the population percentages for founder, m1, m2, and m12 cells were .94, .036, .018, .0004 respectively. All samples are of founder type and we see the star genealogy characteristic of founder lineages.

Finally, Figure 11 shows a genealogy sampled under a constant model and $\lambda_1 = 1, \lambda_2 = .1$. This attack is asymmetric and due to the small value of $\lambda_2$, m12 cells take a long time to sweep through the population. We chose $T_{\text{sample}} = 150$, a time period outside the acute stage, so that we could sample such an asymmetric sweep once m12 cells have fixed. We see few coalescent events. Essentially this asymmetric double attack is equivalent to two single attacks and the result is few coalescent events.

4 Discussion

We use our sampler to explore two aspects of HIV infection. First, we have mentioned on several occasions that due to heavy tails in $D_{\text{stoch}}$, $D_{\text{stoch}}$ often dominates $D_{\text{det}}$. As we have pointed out, these heavy tails can impact the genealogy of infected cell sampled. However, the heavy tails can also change the cell population size dynamics, even for times when such dynamics can be described deterministically. Figures 12 and 13 show population dynamics for the same model parameters: linear model in pre-attack period and CTL attack with rates $\lambda_1 = \lambda_2 = .5$. We see the initial expansion of founder cells and eventually the sweep of m12 cell types in both figures. However, in Figure 12, there is an initial m1 cell birth early in the pre-attack stochastic period. This single initial m1 cell creates a large $D_{\text{stoch}}$ value for the m1 cells allowing them
Table 3: CTL Attack Symmetry and HIV Genetic Diversity

<table>
<thead>
<tr>
<th>$\lambda_1$</th>
<th>0</th>
<th>.05</th>
<th>.10</th>
<th>.15</th>
<th>.20</th>
<th>.25</th>
<th>.30</th>
<th>.35</th>
<th>.40</th>
<th>.45</th>
<th>.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E[L(g)]/(nT_{sample})$</td>
<td>.93</td>
<td>.93</td>
<td>.92</td>
<td>.92</td>
<td>.91</td>
<td>.85</td>
<td>.80</td>
<td>.78</td>
<td>.76</td>
<td>.75</td>
<td>.74</td>
</tr>
</tbody>
</table>

to dominate m2 cells in the interval between founder cell expansion and m12 cell sweep. Figure 13, reflects dynamics in which $D_{stoch}$ was of similar value for both m1 and m2 cells, creating essentially symmetric dynamics.

Second, we use the sampler and our theoretical developments to consider optimal CTL attack. Ideally, CTL attack leads to reduction in infected cell population size. However, if infected cell growth cannot be fully controlled, a secondary goal of CTL attack is the reduction of genetic diversity in the infecting HIV population. Reduced diversity is valuable as it allows for greater efficacy of future CTL attacks since more portions of the HIV genome will be conserved. In our setting, the genetic diversity in question is not epitope diversity, since the e1 and e2 epitopes are already under CTL attack, but rather genetic diversity in other parts of the genome created by neutral mutations. Diversity due to neutral mutations can be explored by forming genealogies for hypothetically sampled cells and placing neutral mutations along the branches according to a given mutation model. Regardless of the mutation model, genealogies with small total branch length will tend to produce less diversity than genealogies with large total branch length. Given a genealogy $g$, let $L(g)$ be the total branch length of $g$. We then consider the values of $\lambda_1, \lambda_2$ that minimize $E[L(g)]$ conditioned on $\lambda_1 + \lambda_2 = w$. We take $w$ to be the total force of CTL attack. Table 3 gives $E[L(g)]/(nT_{sample})$ for $w = 1, n = 20$ and a constant model of pre-attack expansion. We normalize by $nT_{sample}$ since this is the maximum branch length possible. As can be seen, symmetric attack is best at restricting diversity. As mentioned in section 3.3, asymmetric attack can be thought of as two separate single attacks and the comments of section 3.1 show that single attack tends to create few coalescent events.

Our current sampler needs to be extended in several ways. First, the limitation of attack at two epitopes needs to be removed. Our main difficulty in this context is speed of computation. The number of possible infected cell types rises to eight under triple attack and this comes at great computational expense. However, improvements in the C++ implementation should allow for reasonably rapid sampling of attack models with more than two epitopes. This is ongoing work. Second, our model makes several restrictive assumptions that do not hold in typical HIV infection. Namely, the model assumes a fixed infected cell population size after $T_{peak}$, equal fitness of mutant and wild types prior to CTL attack (typically mutations come with fitness costs), and CTL attack at different epitopes that initiate simultaneously. Each of these limitations should not be difficult to remove from a mathematical perspective. However, in each case the main difficulty lies in providing an appropriate model for the more realistic setting. Finally, as mentioned in the introduction, to infer parameter values based on genetic data one needs to associate a probability with a given
genealogy, as opposed to sampling genealogies according to the correct distribution. We feel that our current methods can be modified to achieve this aim, but this too is ongoing work.

Acknowledgements
I thank two anonymous reviewers for many suggestions that greatly improved this paper. I thank one of the anonymous reviewers and Jay Taylor for directing me to helpful references involving the Luria-Delbrück distribution.

References


Figure 2: Distribution of $\frac{D_{\text{stoch}}}{D_{\text{det}}}$ with $A = 100$. The histogram was generated by averaging over 1000 simulations. For each value of $\frac{D_{\text{stoch}}}{D_{\text{det}}}$, the first and second bars represent probability as estimated through simulations and (23) respectively.
Figure 3: Distribution of $\frac{D_{\text{stoch}}}{D_{\text{det}}}$ with $A = 1$. The histogram was generated by averaging over 1000 simulations. The outer graph shows the probability over the full interval $[0, 1]$, while the inner graph shows the probability over the interval $[0, .1]$. For the inner graph, the first bar extends off the scale of the graph.
Figure 4: Distribution $M_{12}(T_{\text{peak}})$. Histogram was generated using 1000 simulations.
Figure 5: Pre-attack genealogy for sampled cells of founder type
Figure 6: Pre-attack genealogy for sampled cells of m1 type
Figure 7: Pre-attack genealogy for sampled cells of m1 type with an early initial m1 cell birth
Figure 8: Symmetric double attack genealogy with cell sampling after m12 sweep
Figure 9: Symmetric double attack genealogy with cell sampling during m12 sweep
Figure 10: Symmetric double attack genealogy with cell sampling at onset of m12 sweep
Figure 11: Asymmetric double attack genealogy
Figure 12: Infected cell population size dynamics with an early initial m1 cell birth
Figure 13: Infected cell population size dynamics without an early initial m1 cell birth